sustainability

MDPI

Article

Deep Learning for Predicting Hydrogen Solubility in n-Alkanes: Enhancing Sustainable Energy Systems

Afshin Tatar, Amin Shokrollahi *, Abbas Zeinijahromi and Manouchehr Haghighi

School of Chemical Engineering, Discipline of Mining and Petroleum Engineering, The University of Adelaide, Adelaide, SA 5005, Australia; afshin.tatar@adelaide.edu.au (A.T.); abbas.zeinijahromi@adelaide.edu.au (A.Z.); manouchehr.haghighi@adelaide.edu.au (M.H.)

* Correspondence: amin.shokrollahi@adelaide.edu.au or shokrollahi.amin@gmail.com

Abstract: As global population growth and urbanisation intensify energy demands, the quest for sustainable energy sources gains paramount importance. Hydrogen (H_2) emerges as a versatile energy carrier, contributing to diverse processes in energy systems, industrial applications, and scientific research. To harness the H_2 potential effectively, a profound grasp of its thermodynamic properties across varied conditions is essential. While field and laboratory measurements offer accuracy, they are resource-intensive. Experimentation involving high-pressure and high-temperature conditions poses risks, rendering precise H_2 solubility determination crucial. This study evaluates the application of Deep Neural Networks (DNNs) for predicting H_2 solubility in n-alkanes. Three DNNs are developed, focusing on model structure and overfitting mitigation. The investigation utilises a comprehensive dataset, employing distinct model structures. Our study successfully demonstrates that the incorporation of dropout layers and batch normalisation within DNNs significantly mitigates overfitting, resulting in robust and accurate predictions of H_2 solubility in n-alkanes. The DNN models developed not only perform comparably to traditional ensemble methods but also offer greater stability across varying training conditions. These advancements are crucial for the safe and efficient design of H_2-based systems, contributing directly to cleaner energy technologies. Understanding H_2 solubility in hydrocarbons can enhance the efficiency of H_2 storage and transportation, facilitating its integration into existing energy systems. This advancement supports the development of cleaner fuels and improves the overall sustainability of energy production, ultimately contributing to a reduction in reliance on fossil fuels and minimising the environmental impact of energy generation.

Keywords: hydrogen solubility; deep learning; machine learning; predictive modelling; sustainable energy

Citation: Tatar, A.; Shokrollahi, A.; Zeinijahromi, A.; Haghighi, M. Deep Learning for Predicting Hydrogen Solubility in n-Alkanes: Enhancing Sustainable Energy Systems. *Sustainability* **2024**, *16*, 7512. https://doi.org/10.3390/su16177512

Academic Editors: Weihua Cai, Chao Xu, Zhonghao Rao, Fuqiang Wang and Ming Gao

Received: 1 August 2024
Revised: 16 August 2024
Accepted: 26 August 2024
Published: 30 August 2024

1. Introduction

The rapid growth of the global population alongside the ongoing trend in urbanisation has resulted in a significant surge in energy requirements. This escalating demand for energy necessitates the exploration of innovative, dependable, and environmentally friendly energy sources. One such solution is the utilisation of hydrogen (H_2) gas as a versatile and sustainable energy carrier [1]. Therefore, it is crucial to possess a comprehensive understanding of the thermodynamic properties of H_2 under various conditions. This knowledge is indispensable for effectively navigating the behaviour of H_2 across diverse pressure (P), temperature (T), and environmental contexts. By delving into the intricate interplay of H_2's thermodynamic characteristics, researchers and engineers can make informed decisions, optimise processes, and ensure the safe and efficient utilisation of H_2 in a wide range of applications. Whether in energy systems, industrial processes, or scientific investigations, a profound grasp of H_2's thermodynamics empowers us to harness its potential with precision and confidence.

H_2 holds substantial significance within the realms of both the petroleum and chemical industries, exemplifying its multifaceted utility. In the pursuit of enhancing the quality of

heavy petroleum fractions, a pivotal strategy involves elevating the H_2-to-carbon ratio. This objective is achieved by incorporating H_2 into hydrocarbons through the hydrocracking process [2]. Consequently, H_2 solubility in hydrocarbon systems emerges as a pivotal thermodynamic parameter, exerting considerable influence over the design, optimisation, and efficiency of diverse chemical and petroleum industrial processes, as well as the associated equipment.

The solubility of H_2 in hydrocarbon systems is influenced by *P*, *T*, and the nature of the hydrocarbon compound. Taking a thermodynamic perspective, the solubility of H_2 in hydrocarbons increases with the increase in *T*, *P*, and the hydrocarbon's Carbon Number (*CN*). This trend has been substantiated by experimental findings documented in the literature [3–5]. An elevated *P* and *T* and a higher *CN* of hydrocarbons foster greater interaction between H_2 molecules and the hydrocarbon matrix, leading to enhanced solubility.

Although field and laboratory measurements of H_2 solubility in hydrocarbons provide precise results, both methods are demanding in terms of time and resources. However, engaging in comprehensive experiments involving heavy hydrocarbon systems under conditions of elevated *P* and *T* introduces a considerable level of risk, rendering this option unappealing within the industry. Consequently, the rapid and accurate determination of H_2 solubility is of utmost importance. In response to these challenges, the industry seeks an approach that efficiently balances accuracy and speed in determining H_2 solubility. Rapid and precise H_2 solubility determination has transformative implications, fostering safe and efficient decision-making within various sectors, including chemical and petroleum industries.

Empirical paradigms, the Equation of States (EoS), and intelligent strategies present promising avenues for predicting H_2 solubility in hydrocarbon systems, offering expedited and cost-effective alternatives to experimental measurements. Nonetheless, the inherent complexity and non-linear nature of H_2 solubility's dependence on *P*, *T*, and the characteristics of n-alkanes complicate the effectiveness of traditional empirical correlations and EoS methods. One of the challenges with EoS methods is the time-consuming process of calibrating various parameters for each specific system. This involves extensive adjustments that can be computationally intensive, particularly when striving for high accuracy across different n-alkanes and operational conditions. Furthermore, near the critical point, where phase behaviour is particularly sensitive, EoS models often face significant challenges in maintaining accuracy. The non-linear interactions between H_2 and hydrocarbons become even more pronounced in these regions, further complicating the prediction process [6–8]. Consequently, the development and application of advanced predictive models, potentially incorporating Machine Learning (ML) techniques, emerge as valuable pursuits in enhancing the accuracy and reliability of H_2 solubility predictions. Such models can better navigate the intricate relationships that underlie H_2 solubility behaviour across diverse hydrocarbon systems and operational conditions. A comprehensive literature review on the mentioned paradigms is provided in our previous study [9].

Recent advancements in ML and deep learning have seen their application in various aspects of renewable energy research, such as optimising the operation of electricity–gas–heat-integrated multi-energy microgrids under uncertainties [10], enhancing security in real-time vehicle-to-grid dispatch [11], improving power forecasting in renewable power plants through novel graph structures [12], calculating dew point pressure in gas condensate reservoirs [13], and the application of Decision Trees (DTs) for the calculation of H_2 solubility in different chemicals [14]. In line with these developments, our study leverages Deep Neural Networks (DNNs) to accurately predict H_2 solubility in n-alkanes, contributing to the efficient design of H_2-based energy systems. This work underscores the growing importance of advanced modelling techniques in promoting sustainable energy solutions.

The primary objective of this study is to evaluate the feasibility of employing DNNs for predicting H_2 solubility in n-alkanes. The investigation focuses on two pivotal aspects. First, we analyse the impact of distinct model structures on predictive performance. Second, we investigate the influence of incorporating dropout layers to mitigate overfitting. To

achieve these goals, three distinct DNN models are constructed, compiled, and trained. The development of these models follows robust methodologies, ensuring the reliability of the results. Extensive assessments are carried out to evaluate the accuracy of each model, ensuring their effectiveness in delivering reliable predictions. In the final stages of this study, a comprehensive stability analysis is executed to assess both the accuracy and precision of the developed model. This analysis is designed to ascertain the generalisability of the developed models. Through this process, we gain valuable insights into the model's performance consistency and its ability to extrapolate knowledge to previously unseen data.

This paper comprises four distinct sections, each serving a specific purpose in addressing the research objectives. It begins with a concise introduction that outlines the context and aims of this study. Following this, Section 2 presents a detailed description of the modelling approaches and the database utilised, providing insights into their composition and characteristics. Section 3 presents the results and discussions. It covers diverse aspects, including the development of predictive models, analysis of errors, evaluation of stability, and a comparison with existing literature models. This section provides a comprehensive understanding of the models' performance and their implications. This paper concludes in Section 4 with a summary of key insights derived from this study's findings and outlines future prospects.

2. Modelling

The initial phase of model development entails data acquisition, a critical foundation for building a robust ML model. The next step involves dividing the database. In this study, the dataset is separated into three sets of training, validation, and testing. Although extensive data cleaning and quality checks were carried out in our previous study [9], here, the database was reviewed for any dubious sample. During model fitting, it is imperative to use only the training and validation sets. The developed model was then applied on the testing sets. The following sections provide detailed discussions on data splitting, model development, and testing data modelling.

The framework illustrated in Figure 1 serves as a roadmap, outlining the sequence of steps integral to the development of the model. As is shown, there are three main steps: data preparation, training (enclosed by blue dashed line), and testing (enclosed by red dashed line). Data preparation includes database development and splitting. The scaler and model are developed in the training phase and are then used in the testing step. During the training phase, it is crucial to utilise only the training and validation subsets. This deliberate isolation is an approach designed to enhance the model's ability to generalise beyond the specific instances on which it has been trained. By restricting the model's exposure to the testing sets, the integrity of the evaluation process is maintained, ensuring that performance assessments remain unaffected by any unintended familiarity with the testing data. Following the model development, the resulting model is subsequently applied to the testing sets. This stage serves as a test of the model's predictive capability and its ability to generalise to unseen data. A thorough evaluation against the testing sets validates the model's real-world applicability and its capacity to provide informed predictions beyond the training context.

In the following sections, a thorough examination was conducted to clarify the complexities of data partitioning, the methodologies used in model development, and the rigorous evaluation of the developed model. Through these discussions, a comprehensive understanding of the challenges and details of the methodology is presented, along with the valuable insights it has the potential to generate.

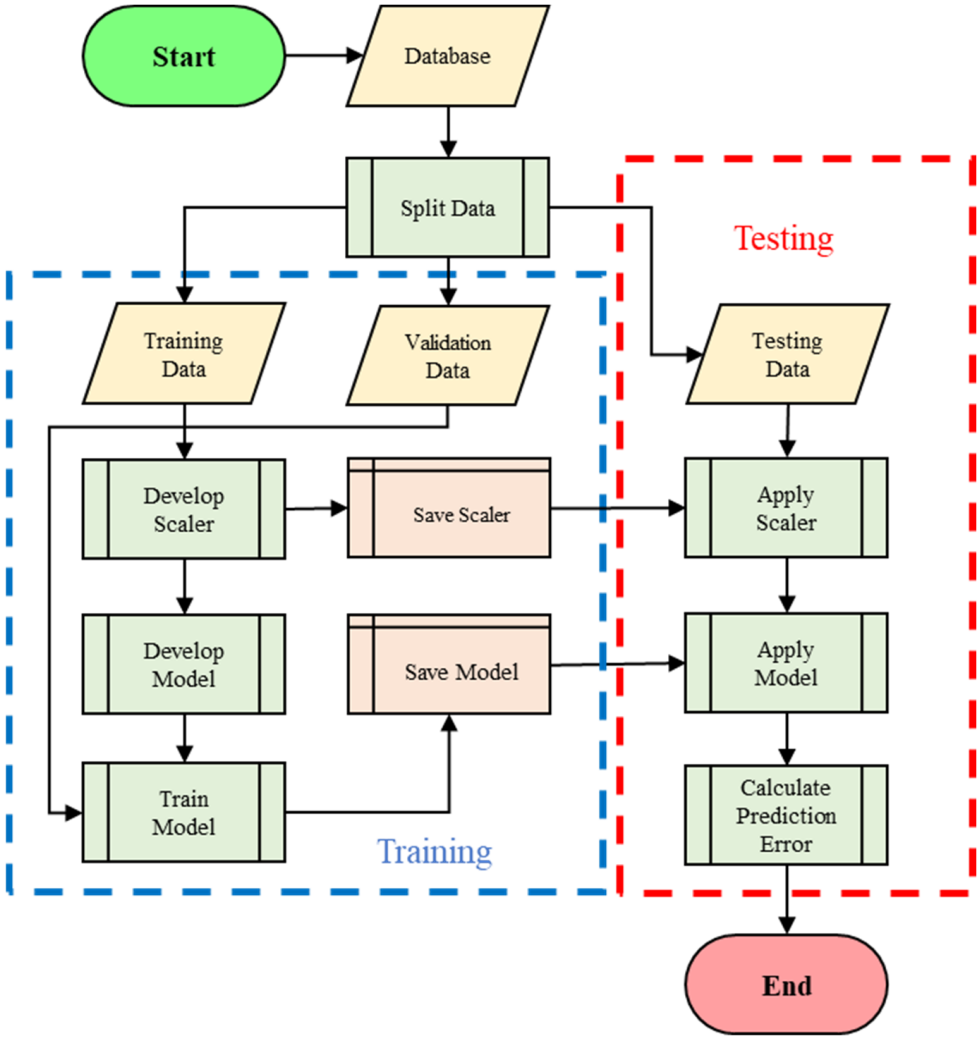

Figure 1. The roadmap for model development used in this study.

2.1. Database Development

A full presentation of the database development is provided in our previous study [9]. All the experimental samples were gathered from the open literature [3,5,15–41]. The database underwent an additional review to ensure data quality. To avoid sparse data samples, we focused on a specific range of pressure (0.101–559.5 MPa) and temperature (92.3–664.05 K). Compared to the previous study [9], the operational variable cut-offs were adjusted, and no Hat-outliers [42,43] were excluded from the database.

The solubility of H_2 in n-alkanes depends on two primary categories of independent variables: the type of n-alkane and operational factors. While a range of characteristics can be used to describe n-alkanes, this study focused specifically on the critical features essential for accurate estimation. The selected critical features for this study include *CN*, critical temperature (T_C) in Kelvin (K), and critical pressure (P_C) in MPa. Operational factors are represented by P and T, which reflect the conditions under which solubility is measured. The characteristics of the n-alkanes utilised in this study are detailed in Table 1.

Furthermore, to enhance the analysis, two engineered features—dimensionless temperature (T_D) and dimensionless pressure (P_D)—are introduced. These features are derived by dividing the actual values of T and P by their respective critical values. Consequently, the modelling process encompasses three types of features: three molecular characteristics (*CN*, T_C, and P_C), two operational variables (T and P), and two engineered features (T_D and P_D). All of these function as independent variables in the predictive model.

Table 1. The characteristics of n-alkanes used in this study [44–50].

n-Alkane	*CN*	P_C (MPa)	T_C (K)
methane	1	4.60	190.56
ethane	2	4.87	305.32
propane	3	4.25	369.83
n-butane	4	3.79	425.12
n-pentane	5	3.37	469.70
n-hexane	6	3.02	507.49
n-heptane	7	2.73	540.13
n-octane	8	2.49	568.88
n-decane	10	2.10	617.70
n-dodecane	12	1.82	658.10
n-hexadecane	16	1.43	722.10
n-eicosane	20	1.20	771.40
n-octacosane	28	0.95	844.00
n-hexatriacontane	36	0.85	896.00
n-hexatetracontane	46	0.45	1064.86

2.2. Data Split to the Training and Testing

In this study, the database was divided into three distinct sets: training, validation, and testing. The model fitting process employs the training data, while the validation dataset was utilised to assess the performance of the trained models during the training phase. Upon successful completion of both training and validation, the model was subsequently tested using data that were not seen during the training phase. In the previous study [9], the performance of these models was evaluated using n-eicosane, comprising 36 samples. To ensure a fair and equitable comparison, the same methodology was adopted in this study.

To enhance the robustness of the results, data splitting in this study was conducted based on n-alkanes rather than individual data samples. Specifically, the division was performed on the count of n-alkanes, ensuring that all samples associated with a particular chemical were consistently assigned to the same dataset. This methodology encourages the model to independently learn the complexities of developing isotherms, rather than simply focusing on the task of imputing missing data points. Figure 2 provides a clear illustration of the partitioning of data into training and testing sets for three distinct chemicals within a fixed *P*. Figure 2a illustrates the sample-wise data division, while Figure 2b depicts the group-wise division. It is worth noting that in sample-wise splitting, the testing data points are combined with the training data, facilitating the potential for predicting testing data through interpolation. In contrast, group-wise splitting assigns the testing data to a chemical that is not represented in the training set. Essentially, this method necessitates that the model understands and predicts underlying trends based on the distinct characteristics of each chemical.

Table 2 presents the various sets, detailing the names of the n-alkanes alongside the corresponding sample counts for each set. Of the 15 n-alkanes, 9 are assigned to the training set, 3 to the validation set, and 3 to the testing set, resulting in a nominal data split ratio of 60:20:20. However, owing to the differing sample counts for the various n-alkanes, the actual split ratio based on sample count is approximately 76:13:11. This discrepancy arises primarily from the relatively high number of methane samples (297) included in the training set. It is noteworthy that both the validation and testing sets comprise n-alkanes that were not part of the training phase, thereby ensuring a robust evaluation of the model.

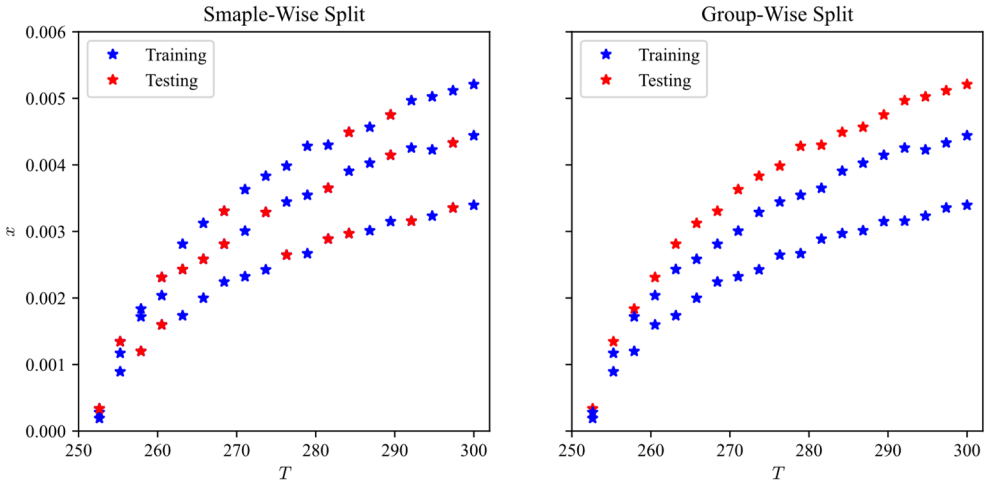

Figure 2. Comparison of (**a**) sample-wise and (**b**) group-wise data splitting.

Table 2. The data split into the training and testing based on the n-alkane type.

Index	n-Alkane	*CN*	*N*
	Training (*N* = 1163)		
1	methane	1	297
2	propane	3	102
3	n-butane	4	92
4	n-heptane	7	13
5	n-octane	8	70
6	n-decane	10	247
7	n-hexadecane	16	172
8	n-octacosane	28	108
9	n-hexatetracontane	46	35
	Validation (*N* = 196)		
1	ethane	2	61
1	n-pentane	5	105
3	n-dodecane	12	24
	Testing (*N* = 162)		
2	n-hexane	6	56
2	n-eicosane	20	36
3	n-hexatriacontane	36	66

2.3. Input Preparation

As previously indicated, the steps of data cleaning and quality assessment, including the exclusion of duplicates, feature extraction, and extreme P and T values, were executed in our preceding study [9]. In this study, additional measures were taken: the operational parameter cut-offs were adjusted to remove sparse samples, outliers identified by the Hat-method [42,43,51] were included, and the database underwent another thorough review to exclude any dubious samples.

Constructing ML models using scaled data is regarded as a sound practise. In this study, standardisation was employed as the selected scaling technique. This process involves subtracting the mean value of a feature from each individual feature value and then dividing the resultant value by the standard deviation of that feature. As a result of this transformation, the feature achieves a mean of 0 and a standard deviation of 1, facilitating consistent and standardised comparisons between different features.

While scaling may not be obligatory for non-parametric models like DT-based models due to their inherent insensitivity to feature scaling, its significance becomes pronounced for distance-centric models such as the DNN and Support Vector Machine (SVM). These models significantly depend on the distance metrics between data points, and the presence of unscaled features may distort these distance computations, adversely affecting the model's performance and convergence. Furthermore, employing scaled data generally results in reduced computational time during the modelling phase. When input features are normalised to a similar scale, the convergence of algorithms can be expedited, facilitating quicker optimisation. Additionally, using scaled data often contributes to a more stable training process, as it mitigates the risk of features with large values to dominate the learning process.

The characteristics of the training, validation, and testing data are provided in Table 3. As shown, the skewness and kurtosis of the operational parameters are close to zero. This indicates that their distribution is close to normality, suggesting a balanced dataset without significant outliers or extreme values. A near-zero skewness implies a symmetric distribution of the data around the mean, while a near-zero kurtosis indicates that the data's tails are not heavy, thus reducing the likelihood of anomalies. This balance in the dataset enhances the reliability and accuracy of the model's predictions. Another important point is that the *CN* is not considered a categorical feature. Considering *CN* as a categorical variable would limit the model to the n-alkanes encountered during the training phase, which is not desirable. Our goal is to develop a model applicable to all possible n-alkanes, including those not used for training, ensuring broader applicability and robustness.

Table 3. The statistical characteristics of the variables used in this study for the training, validation, and testing sets.

Set	Parameter	*CN*	T_C (K)	P_C (MPa)	T (K)	P (MPa)	T_D	P_D	x
Training	*Min*	1	190.56	0.45	92.30	0.65	0.27	0.14	0.0021
	Q_1	1	190.56	1.43	173.15	4.04	0.52	1.50	0.0398
	Median	8	568.88	2.49	344.30	6.74	0.63	2.94	0.0762
	Q_3	16	722.10	4.60	423.20	10.60	0.81	4.71	0.1233
	Max	46	1064.86	4.60	583.45	28.96	0.98	35.18	0.5013
	Mean	10.14	515.80	2.85	320.95	8.19	0.66	3.72	0.0951
	SD	10.27	242.87	1.42	136.85	5.84	0.17	3.66	0.0794
	IQR	15	531.54	3.17	250.05	6.56	0.29	3.21	0.0835
	skewness	1.62	0.05	0.03	−0.16	1.29	0.11	3.54	1.8771
	kurtosis	2.69	−0.93	−1.57	−1.13	1.43	−1.08	19.27	4.5143
Validation	*Min*	2	305.32	1.82	92.50	0.69	0.30	0.21	0.0044
	Q_1	2	305.32	3.37	228.15	4.80	0.58	1.35	0.0259
	Median	5	469.70	3.37	338.15	7.24	0.71	2.05	0.0531
	Q_3	5	469.70	4.87	383.15	11.24	0.82	3.29	0.0960
	Max	12	658.10	4.87	463.15	29.57	0.99	8.20	0.2917
	Mean	4.89	439.93	3.66	310.40	9.00	0.70	2.54	0.0663
	SD	2.98	110.21	0.97	102.61	6.32	0.17	1.68	0.0473
	IQR	3	164.38	1.51	155.00	6.44	0.23	1.94	0.0700
	skewness	1.38	0.38	−0.23	−0.57	1.41	−0.46	1.14	1.3991
	kurtosis	1.31	−0.38	−0.57	−0.73	1.66	−0.15	1.07	3.3077
Testing	*Min*	6	507.49	0.85	298.15	0.99	0.40	0.41	0.0105
	Q_1	6	507.49	0.85	372.52	3.01	0.45	1.83	0.0355
	Median	20	771.40	1.20	377.60	5.07	0.53	3.37	0.0676
	Q_3	36	896.00	3.02	423.20	8.24	0.68	6.00	0.1114
	Max	36	896.00	3.02	573.25	16.75	0.81	19.82	0.2271
	Mean	21.79	730.84	1.69	395.06	5.76	0.57	4.68	0.0795
	SD	13.16	171.40	0.99	62.98	3.56	0.13	4.01	0.0524
	IQR	30	388.51	2.17	50.68	5.23	0.22	4.17	0.0760
	skewness	−0.08	−0.40	0.57	1.16	0.73	0.46	1.45	0.9179
	kurtosis	−1.72	−1.64	−1.62	1.74	−0.18	−1.16	1.70	0.2960

2.4. Model Development

The primary objective of this study is to evaluate the effectiveness of DNNs in predicting H_2 solubility across a range of n-alkanes. To achieve this aim, three distinct DNNs were developed. The varied architectural compositions of these models offer a comprehensive framework for examining the effects of incorporating batch normalisation [52] and dropout layers [53], as well as variations in layer arrangement. Batch normalisation enhances the training speed and stability of DNNs, while dropout mitigates overfitting by randomly omitting units and their connections during the training process. It is important to highlight that this study utilised Python, along with Keras running on the TensorFlow backend, for the modelling process. A list of all the packages employed, along with their respective versions and the specifications of the computer system used for modelling, can be found in Appendix A.

2.4.1. Model Construction

Keras was utilised to construct the models. Considering the necessity of evaluating layer concatenation, the functional Application Programming Interface (API) was selected over the simpler sequential API. Three models, designated as DNN 1, DNN 2, and DNN 3, were examined, with the details of these models outlined in the subsequent section.

DNN 1 represents the most straightforward model under consideration. As depicted in Figure 3a, this model consists of three hidden layers, each containing 30 neurons.

DNN 2 represents an enhanced version of DNN 1, achieved by integrating batch normalisation and dropout into every hidden layer, as illustrated in Figure 3b. There are discrepancies among researchers regarding the nomenclature of these layers; some classify DNN 2 as a 10-layer network, comprising nine hidden layers and one output layer. In this study, as shown in Figure 3, we adopted the term "block" to refer to a unit that encompasses the primary layer (Dense) along with its associated components (batch normalisation and dropout). To aid clarity, distinct colours were assigned to each type of layer.

The configuration of DNN 3 is illustrated in Figure 4. Similar to DNN 2, it comprises three blocks consisting of dense layers, batch normalisation, and dropout layers. However, the first block exhibits a notable distinction. As previously mentioned, the target variable depends on three primary inputs, P and T, which represent the operational parameters, along with the type of n-alkane. Furthermore, two additional features, P_D and T_D, were derived by integrating operational and molecular characteristic attributes.

As illustrated in Figure 4, the network inputs are categorised into three segments: Input 1, Input 2, and Input 3. These segments represent molecular characteristics (comprising three features), engineered features (encompassing two features), and operational features (incorporating two features), respectively. Each segment is connected to a hidden layer consisting of ten units. Following the processes of batch normalisation and dropout, these segments are concatenated to create a layer comprising thirty units. The subsequent architecture is consistent with that of DNN 2.

To provide a more comprehensive insight into the model's structure, a dropout ratio of 0.05 was employed, meaning that approximately 5% of the neurons were temporarily excluded during training. This approach enhances generalisation and mitigates the risk of overfitting. The Adam optimiser was selected to compile the models, which is a standard practise in model optimisation. By iteratively adjusting the model's parameters, the optimiser minimises the influence of the chosen loss function. In this case, the loss function was defined as the Mean Squared Error (*MSE*), which is a suitable choice for regression tasks, quantifying the average squared difference between predicted and actual values.

The parameter count, which includes both trainable and non-trainable parameters, is thoroughly detailed in Table 4. This count has a direct impact on the complexity of the model and its potential performance. Notably, the inclusion of batch normalisation adds both trainable and non-trainable parameters to the model. This technique not only stabilises and accelerates the training process but also enhances the overall performance of the model.

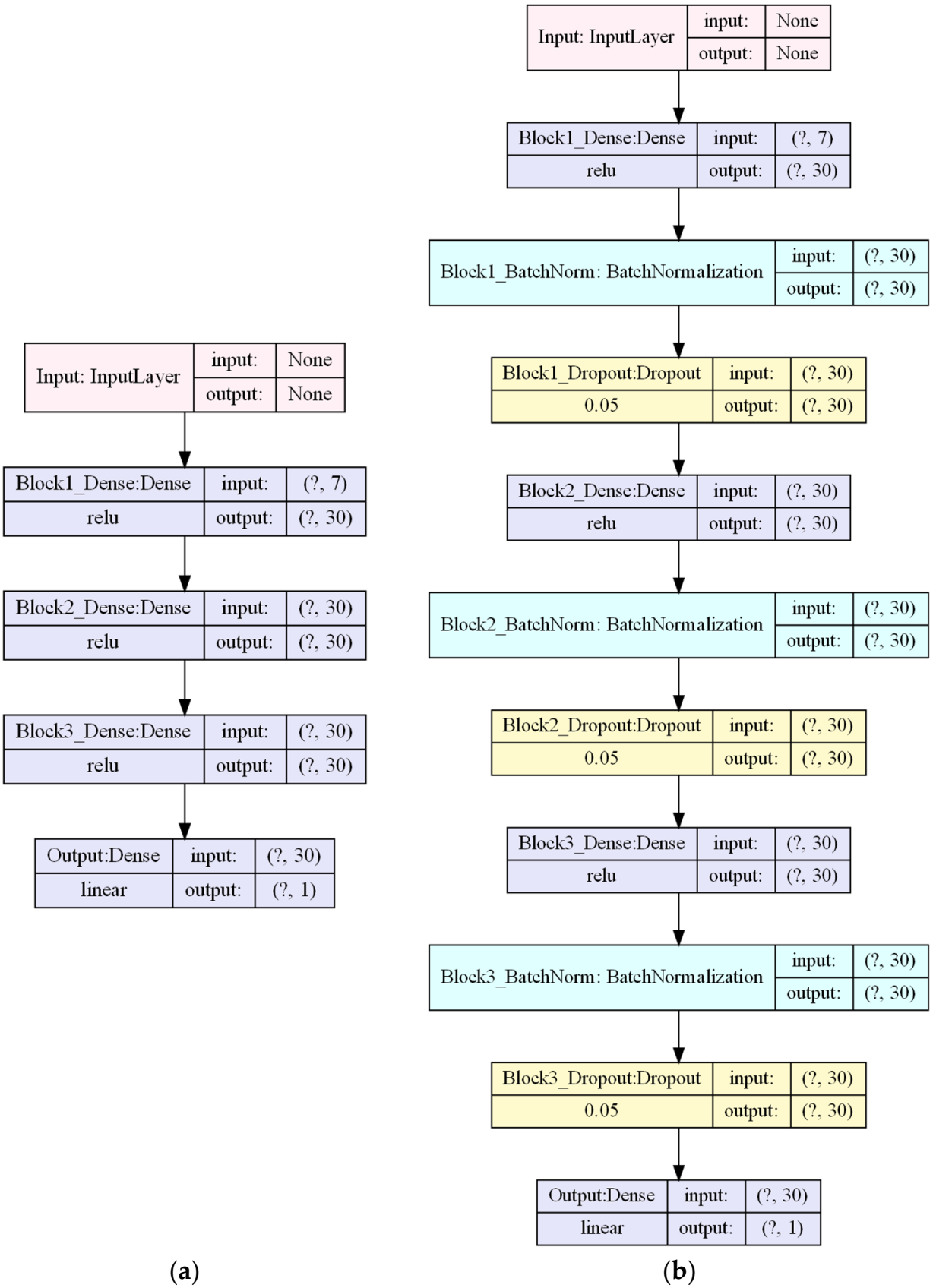

Figure 3. The schematic presentation of (**a**) DNN 1, which has 3 hidden layers, and (**b**) DNN 2, which has 3 blocks for the hidden layers.

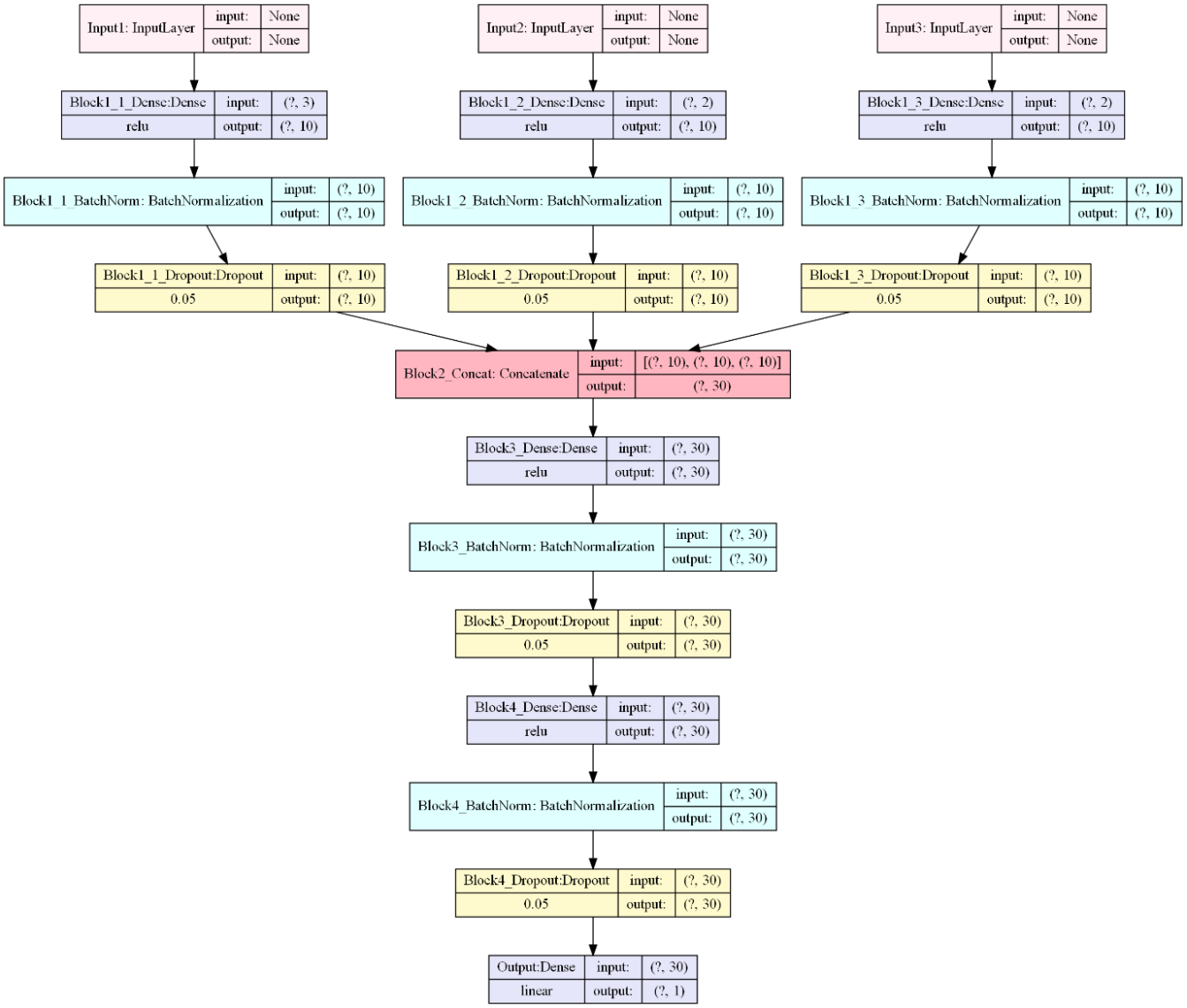

Figure 4. The schematic presentation of DNN 3.

Table 4. Number of parameters in constructed models.

Model	Trainable	Non-Trainable	Total
DNN 1	2131	0	2131
DNN 2	2311	180	2491
DNN 3	2171	180	2351

Notably, DNN 3 stands out by featuring fewer trainable parameters compared to its counterpart, DNN 2. This reduction results from the lack of interconnections between the various input types in its input layer. This streamlined architecture not only diminishes the overall complexity of the model but also aligns effectively with the specific modelling objectives.

2.4.2. Model Training

After constructing and compiling the models, the subsequent phase entails fitting them to the data. During this stage, the models are trained using the provided dataset, with the number of "epochs" and the "batch size" playing crucial roles. Specifically, "epochs" refer to the number of times the entire dataset is iterated over during the training phase,

while 'batch size' determines the number of data points processed before the model's parameters are updated. In this study, the models were trained for 1000 epochs with a batch size of 64.

A critical aspect of this process is monitoring the "validation loss". This metric provides valuable insights into the model's performance on unseen validation data, helping to ensure that the model does not become excessively tailored to the training data and retains its ability to generalise to new information. The purpose of tracking the validation loss is to identify the point at which the model's performance on the validation dataset is optimised. Once this optimal performance stage is reached, the model's configuration is saved as the best iteration using a callback. This "best model" configuration then serves as a reference for future applications and comparisons, ensuring that the most effective model iteration is preserved.

Figure 5 visually illustrates the convergence of the loss function, represented by the *MSE*, for both the training and validation datasets. This representation indicates the extent to which the model's predictions align with the actual data points, providing insight into its predictive effectiveness.

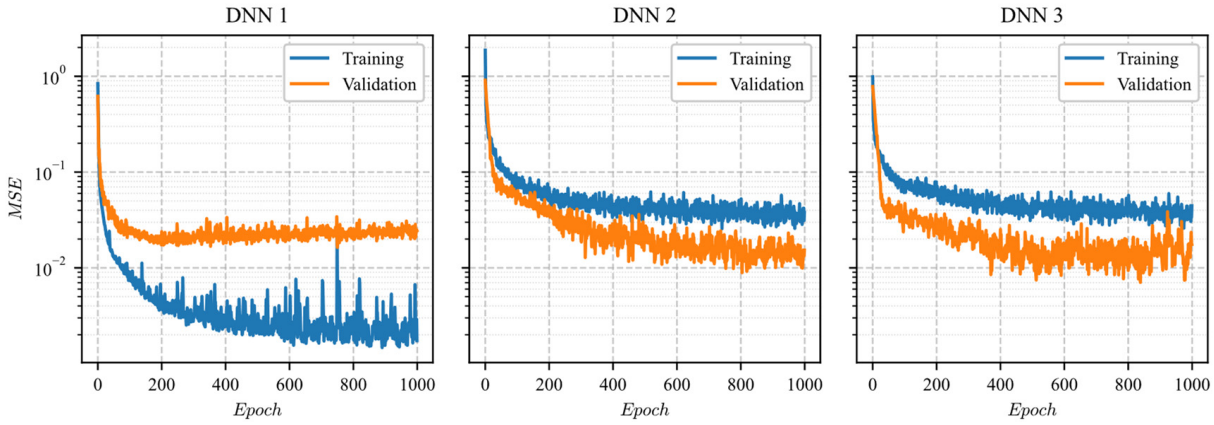

Figure 5. The convergence of the loss function (*MSE*) for the three DNN models.

Examining DNN 1 reveals that, although the training error decreases, there is no corresponding improvement in the testing error. Even after 200 epochs, the validation error exhibits a slight upward trend. This phenomenon, referred to as overfitting, suggests that the model has become excessively tailored to the training data, which compromises its ability to generalise to new, unseen data points.

To address the issue of overfitting, dropout—a technique that temporarily deactivates a subset of neurons during training—was judiciously employed. The implementation of dropout helps mitigate overfitting by improving the model's capacity to generalise beyond the training data. When comparing training losses, DNN 2 and DNN 3 demonstrate higher values than DNN 1. However, both DNN 2 and DNN 3 show a significant reduction in validation loss without raising concerns about overfitting.

A notable distinction emerges when comparing DNN 2 and DNN 3. DNN 3 demonstrates superior performance with respect to validation data, highlighting its enhanced capability to capture underlying patterns within the data. This improved performance contributes to better generalisation on unseen samples.

2.4.3. Predicting the Testing Data

Upon successfully training the models and identifying the best-performing one based on validation loss, the next step involves applying this model to the test data. This process allows for the evaluation of the model's predictive performance on previously unseen data points.

Before inputting the testing data into the network, it is crucial to apply scaling to the data. Additionally, the architecture of the models relies on a logarithmic transformation of the target variable. Once the model generates predictions, an inverse transformation is performed to revert the solubility values to their original scale. This process consists of two main steps: first, the inverse scaling procedure is carried out to reverse the initial data scaling; second, the 10th power is applied to reverse the logarithmic transformation. This results in the predicted solubilities being expressed on their original scale.

3. Result and Discussion

The dependent variable (x) is closely linked to three independent variables: P, T, and the specific chemical type. Together, these independent factors influence the target variable x. To characterise the chemicals under consideration comprehensively, a variety of descriptors can be applied. Each descriptor adheres to its own distinct statistical distribution, highlighting the limitation of relying on a single descriptor. Therefore, exploring multiple descriptors is essential for a more accurate understanding. In addition to the three primary characteristics, two engineered dimensions, P_D and T_D, are introduced. These engineered variables provide a standardised framework for incorporating P and T, facilitating a more cohesive analysis.

To model the target variable x, representing the mole fraction of H_2, a logarithmic transformation of the original data is selected. This approach is informed by a significant observation: the distribution of x exhibits a lognormal pattern, with values predominantly clustering around zero (see Figure 6). The logarithmic transformation serves two key purposes. Firstly, it aids in the development of a normal distribution, which is a common assumption in statistical modelling. Secondly, and perhaps more critically, it prevents the generation of negative predictions for values close to zero. This consideration is vital to ensure that the model's predictions remain consistent with the physical constraints of the data.

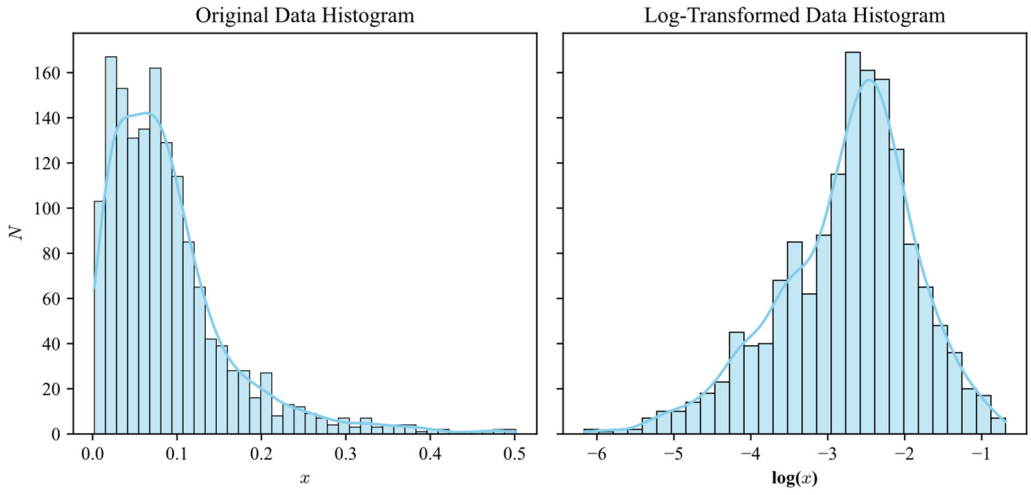

Figure 6. The distribution of target values.

It is noteworthy that DNN models, in contrast to their DT counterparts, possess a unique capability for extrapolation. This allows DNN models to generate predictions that extend beyond the predefined range of target values. Consequently, this feature enhances the model's versatility and its capacity to offer insights into scenarios that fall outside the range of the training data.

3.1. Statistical Error Analyses

This section presents a comprehensive assessment of each model's performance, employing both graphical illustrations and statistical methods. The previously mentioned *MSE*, calculated using logarithmically transformed and scaled solubility values, is not

utilised. Instead, the evaluation focuses on calculating the Root-Mean-Squared Error (*RMSE*) related to the actual solubility values expressed in mole fraction units. This adjustment facilitates a more direct and accessible understanding of the error scale. Additionally, the Symmetric Mean Absolute Percentage Error (*SMAPE*) is calculated, which ranges from 0 to 100%. The formulations for the model metrics utilised are provided in Table 5.

Table 5. Model metrics for assessing the accuracy of developed models.

Metric	Formula	Range	Ideal Value
Residual (Res_i)	$Res_i = y_i - t_i$	$(-\infty, \infty)$	0
Root-Mean-Squared Error (*RMSE*)	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Res_i)^2}$	$[0, \infty)$	0
Symmetric Percentage Error (*SPE*)	$SPE = 100 \times \frac{Res_i}{\lvert y_i\rvert+\lvert t_i\rvert}$	$[-100, 100]$	0
Symmetric Mean Absolute Percentage Error (*SMAPE*)	$SMAPE = \frac{100}{n} \times \sum_{i=1}^{n} \frac{\lvert Res_i\rvert}{\lvert y_i\rvert + \lvert t_i\rvert}$	$[0, 100]$	0

Table 6 provides a comprehensive overview of the model metric values derived from the models developed in this study, each evaluated across distinct datasets. The use of these model metrics offers a quantitative perspective for assessing the performance of the models under various conditions. The top performer in each dataset—training, validation, and testing—is highlighted using bold formatting, which improves the clarity of their identification.

Table 6. Model metrics values for the models developed in this study for different sets.

Model	Set	R^2	*RMSE*	*SMAPE* (%)	*N*
DNN 1	Training	**0.99**	**0.007**	**1.82**	1163
	Validation	0.90	0.015	4.43	196
	Testing	0.93	0.014	6.70	162
DNN 2	Training	0.99	0.009	2.26	1163
	Validation	0.99	0.005	2.96	196
	Testing	**0.98**	**0.007**	**3.24**	162
DNN 3	Training	0.98	0.012	2.64	1163
	Validation	**0.99**	**0.004**	**2.58**	196
	Testing	0.97	0.010	3.28	162

Upon thorough evaluation, DNN 1 is the best model based on its performance on the training set, achieving an *RMSE* of 0.006991 and an *SMAPE* of 1.82%. However, its effectiveness appears to diminish when applied to the validation and testing datasets, as indicated by *RMSE* values of 0.014867 and 0.014058, and *SMAPE* values of 4.43% and 6.70%, respectively. In contrast, DNN 2 and DNN 3 display a notable consistency in their ability to generalise beyond the training data. Both models demonstrate similar error rates in the training and testing sets. Specifically, DNN 2 has a testing set *RMSE* of 0.007050 and an *SMAPE* of 3.24%, while DNN 3 shows a testing set *RMSE* of 0.009641 and an *SMAPE* of 3.28%. Remarkably, DNN 3 stands out for its superior predictive accuracy on the validation sets, outperforming its peers in this regard.

Table 7 presents the model metrics associated with DNN 3 across each n-alkane within the training, validation, and testing sets, providing a detailed view of the model's predictive accuracy. The Symmetric Mean Absolute Percentage Error (*SMAPE*) for all n-alkanes ranges from 1.29% to 4.94% in the validation and testing sets. This relatively narrow error margin across diverse n-alkanes indicates that DNN 3 is highly effective at generalising from the training data to unseen data, maintaining a high level of accuracy even when predicting the solubility of n-alkanes not included in the model's training phase.

Table 7. The model metrics for DNN 3 for each n-alkane.

Index	n-Alkane	R^2	*RMSE*	*SMAPE* (%)	*N*
Training (*N* = 1163)					
1	n-Butane	0.982	0.008	2.62	100
2	n-Decane	0.988	0.007	1.55	253
3	n-Heptane	0.977	0.005	2.27	5
4	n-Hexadecane	0.991	0.004	1.35	181
5	n-Hexatetracontane	0.959	0.013	2.57	36
6	Methane	0.972	0.018	3.86	305
7	n-Octacosane	0.971	0.007	2.78	111
8	n-Octane	0.992	0.005	2.63	70
9	Propane	0.963	0.017	3.87	102
Validation (*N* = 196)					
1	n-Dodecane	0.996	0.002	1.29	24
2	n-Ethane	0.985	0.007	4.51	63
3	n-Pentane	0.996	0.003	1.74	109
Testing (*N* = 162)					
1	n-Eicosane	0.984	0.004	2.40	37
2	n-Hexane	0.996	0.002	1.86	57
3	n-Hexatriacontane	0.939	0.014	4.94	68

The consistency of low *SMAPE* values across different n-alkanes suggests that the model has not only captured the underlying physical relationships governing H_2 solubility but also generalised these relationships well to new data. This ability to generalise is crucial for the practical application of the model in real-world scenarios, where it may need to predict solubility for n-alkanes beyond those included in the initial dataset. Essentially, the DNN 3 model's performance metrics underscore its robustness and reliability, demonstrating that it has effectively learned the governing physical patterns of H_2 solubility in n-alkanes. This strong performance supports the model's potential use in various industrial applications, where accurate and reliable solubility predictions are essential for optimising H_2-based processes and systems.

Figure 7 presents a scatter plot that juxtaposes predicted values against actual experimental values in the upper section, while the lower section depicts the alignment of the Standard Prediction Error (*SPE*) with the experimental values. These values are derived from the validation and testing sets, predicted using the DNN 3 model. To facilitate the comparison of data samples, both plots share a common *x*-axis, ensuring a coherent alignment between the upper and lower sections. A closer examination of the scatter plot reveals that the majority of data points are situated near the 45-degree line, indicating a strong correlation between the model's predictions and the actual experimental values. Additionally, the *SPE* plot demonstrates that most data samples exhibit *SPE* values constrained within −10% and 10%. This figure demonstrates the model's exceptional performance when tested with unseen n-alkanes, indicating that it has effectively identified the fundamental physical patterns and key relationships governing their behaviour. Its ability to predict the behaviour of new n-alkanes not included in the training dataset confirms its capacity to generalise beyond the training data. This robustness highlights the model's potential for practical applications across various scenarios involving n-alkanes, showcasing its capability to provide valuable insights in relevant fields.

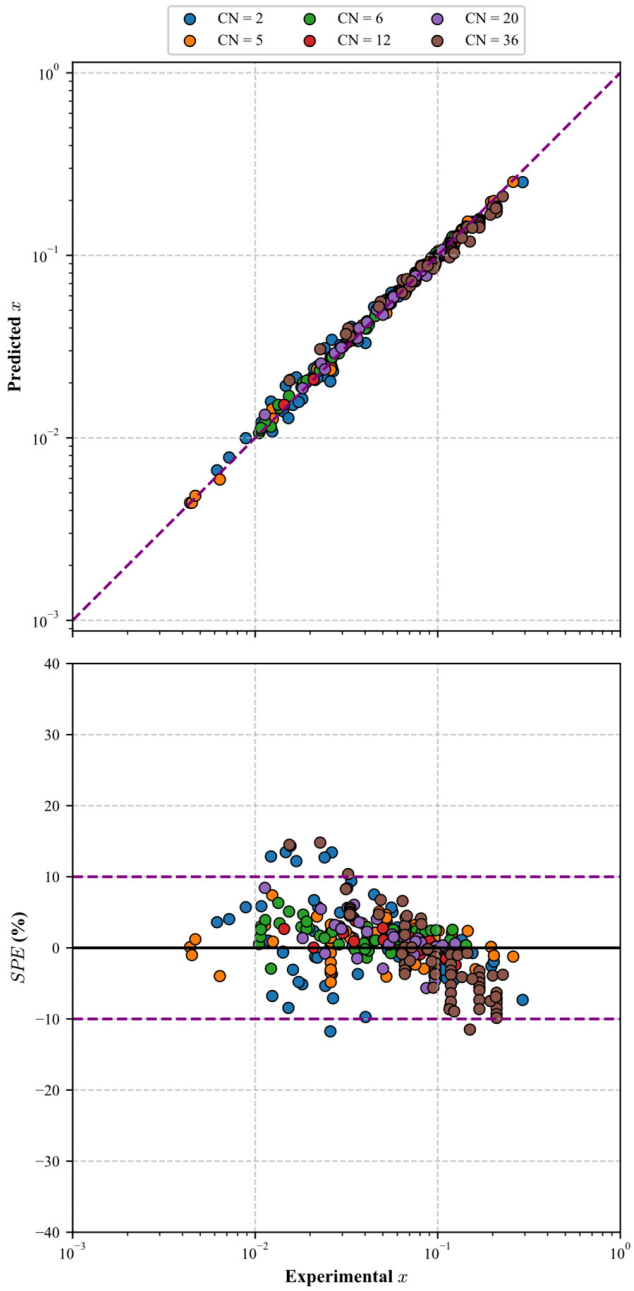

Figure 7. The scatter plot for predicted versus experimental x and SPE versus the experimental x for the validation, testing, and extra testing data predictions yielded by the DNN 3 model.

3.2. Comparison with the Literature Models

In our previous study [9], we developed and tested three DT-based models. These included a basic DT model and three ensemble models: Gradient Boosting (GB), Random Forest (RF), and Extra Trees (ET). Notably, ensemble models aggregate multiple simple DT models, with each employing distinct aggregation techniques. The ensemble models from our prior research utilised a considerable number of simple estimators, specifically incorporating 84 estimators for the GB model, 70 for the RF model, and 90 for the ET model.

The current study introduced a more robust method for data separation, enhancing data quality. However, to ensure equitable comparison, n-eicosane samples were used for extra testing. These data were not used during training of the model. Illustrated in Figure 8 is a cumulative distribution function plot, depicting the absolute *SPE* for n-eicosane. This plot illustrates the DNN models' superior performance compared to both the basic DT

model and the ensemble ET model. Remarkably, the DNN models demonstrate superior efficacy to the GB and RF models.

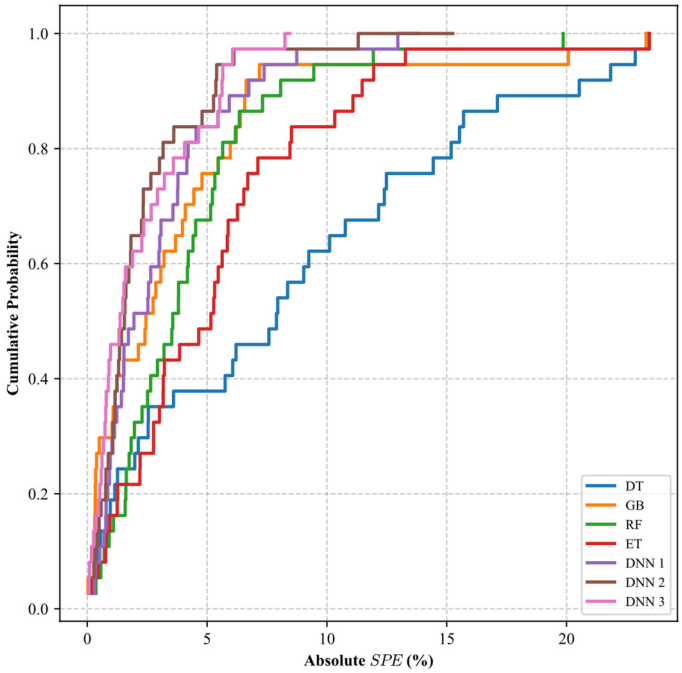

Figure 8. Cumulative distribution function plot for absolute *SPE* (%).

This study presents a more robust method for data separation, which significantly enhances data quality. To ensure a fair comparison, similar to our prior study [9], extra testing was conducted using n-eicosane samples, which were not included during the training of the model. Figure 8 illustrates the cumulative distribution function plot, depicting the absolute *SPE* for n-eicosane. This plot clearly demonstrates the superior performance of the DNN models in comparison to both the basic DT model and the ensemble ET model. Notably, the DNN models also exhibit greater efficacy than the GB and RF models.

Additionally, Table 8 provides model metrics for both our previous study [9] and the DNN models developed in the current research. Among the models published in our earlier work [9], only the RF predictions exhibit a close alignment with the DNN models, with all maintaining an *SMAPE* of less than 5%. Furthermore, the generalisability observed in the DNN models may be attributed to the robust data separation methodology employed in this study.

Table 8. Model metrics for literature models and the DNN models developed in this study.

Model	*RMSE*	*SMAPE* (%)
DT [9]	0.0112	5.34
GB [9]	0.0068	5.15
RF [9]	0.0050	4.02
ET [9]	0.0071	5.44
DNN 1	0.0044	3.27
DNN 2	0.0035	2.63
DNN 3	0.0041	2.40

Nevertheless, it is important to note that the cut-off values for operational parameters were adjusted in this study, and several incorrectly recorded data points were either excluded or corrected. Consequently, the comparison may not definitively demonstrate that the DNN is superior to ensemble DT-based models. Rather, it highlights that the

models developed in this study represent a significant advancement towards achieving greater accuracy and reliability.

3.3. Model Stability

The development of a DNN model involves various elements that introduce a degree of uncertainty. This investigation focuses on two primary factors contributing to this uncertainty. The first factor arises from the initial randomisation of the model's weights, while the second pertains to the random partitioning of data into training, validation, and testing sets. To ensure the model's effectiveness, it must be capable of effectively managing and adapting to these inherent sources of randomness.

To conduct a comprehensive investigation into the effects of stochastic model training and the initialization of models with random weights, a rigorous procedure was established that involved the creation, compilation, and fitting of 50 networks. Particular emphasis was placed on minimising other sources of randomness throughout the experimental process. A key aspect of the methodology was the use of identical datasets, which ensured consistency across the various phases of training and evaluation. The hyperparameters detailed in Section 2.4 were consistently applied during both the compilation and fitting of the models. However, to optimise computation time, all models were trained for 600 epochs instead of the originally planned 1000. While this adjustment may result in a slight reduction in prediction accuracy compared to previous sections, it effectively illustrates the impact of randomness.

Upon completing each iteration of model training, a rigorous testing phase was conducted using the designated testing data. The evaluation metric employed was the *SMAPE*. Figure 9 provides a visual representation of this process, depicting the *SMAPE* values obtained from a diverse set of 50 DNNs, each subjected to distinct training processes. This figure includes a histogram (subplot (a)) and a QQ plot (subplot (b)). Notably, both graphical representations collectively support the conclusion of a normal distribution of the errors.

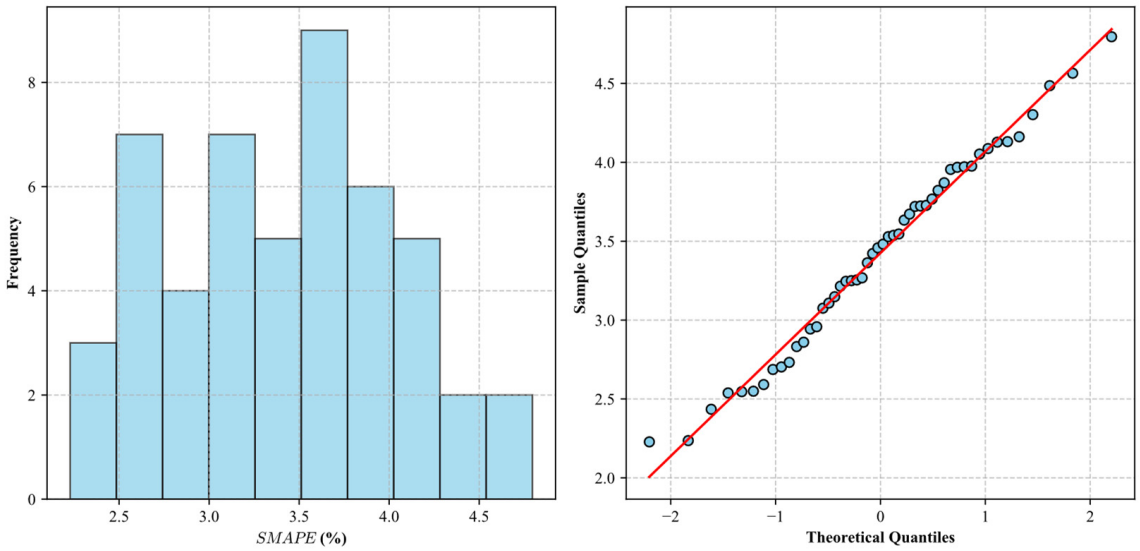

Figure 9. *SMAPE* for 100 DNNs with different training processes: (**a**) distribution reflected in histogram and (**b**) the QQ plot.

The normal distribution of errors across the 50 distinct DNNs offers valuable insights into the model's stability and robustness. This distribution indicates a consistent performance across various training processes, suggesting that the model's behaviour is not unduly affected by random factors, which results in predictable outcomes. Furthermore,

models exhibiting normally distributed errors tend to demonstrate greater robustness, as they are resilient to variations in training conditions. This resilience significantly enhances their ability to generalise effectively to new, unseen data.

In contrast, the subsequent experiment revealed a more pronounced influence of randomness arising from the data partitioning process. A systematic approach was employed for data partitioning, beginning with the segregation of the data samples belonging to n-eicosane, which was designated as the additional testing set in the previous study [9]. These samples were set aside for testing across all models. The remaining dataset was then divided into three distinct subsets—training, validation, and testing—in a ratio of 60:20:20. This division was executed using a group-based methodology aimed at preserving the integrity and coherence of data groups throughout the modelling process.

Figure 10 illustrates both the associated histogram (shown in subplot (a)) and the QQ plot (displayed in subplot (b)). Unlike the first experiment, where the errors exhibited a well-defined normal distribution, the errors in the second experiment displayed a distribution that deviated from the normal pattern. This observation suggests that the randomness introduced by data partitioning had a more substantial impact on the model's performance than the randomness introduced by weight initialization. Consequently, inconsistent data partitioning can lead to increased variability in the model's performance, hindering its ability to generalise effectively to new data. The notable degree of randomness observed in the second experiment can be attributed to the differing distribution of data across the various sets.

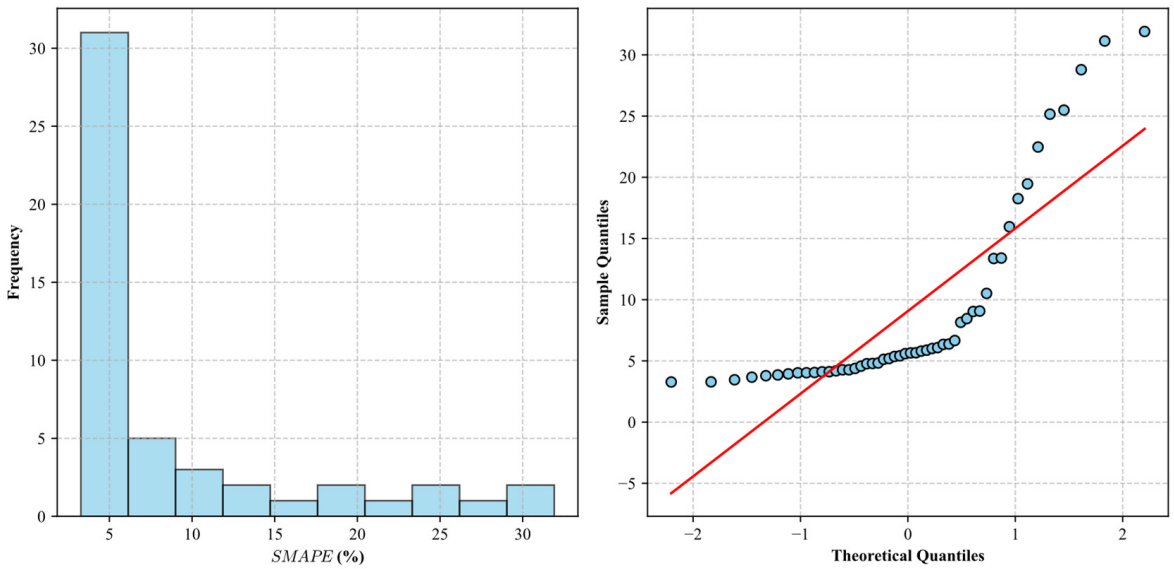

Figure 10. SMAPE for 100 DNNs with different splitting: (**a**) distribution reflected in histogram and (**b**) the QQ plot.

This study's findings underscore the importance of appropriate data partitioning, especially in scenarios where available experimental data samples are limited. In such instances, achieving a consistent distribution of data across different sets is crucial for minimising the adverse effects of randomness on the model's performance. Notably, when a more extensive dataset is available for training, the potential impact of randomness introduced by data partitioning may be reduced, owing to the larger sample size. This observation further highlights the significance of strategic data management, which can ultimately lead to more reliable and robust model outcomes.

Figure 11 provides a graphical representation of the outcomes derived from the two experiments conducted: the training randomness experiment and the splitting randomness

experiment. In this visual depiction, the x-axis and y-axis represent the logarithmically transformed solubility values and P, respectively, for n-eicosane. To facilitate a deeper understanding, the instances were arranged in isotherms. The key observation from this figure is the marked contrast in model performance across the various training trials, which employed a fixed data partitioning approach and different data partitioning methods. Notably, the trials using the fixed approach demonstrate superior accuracy and precision compared to those constructed with different data partitioning strategies.

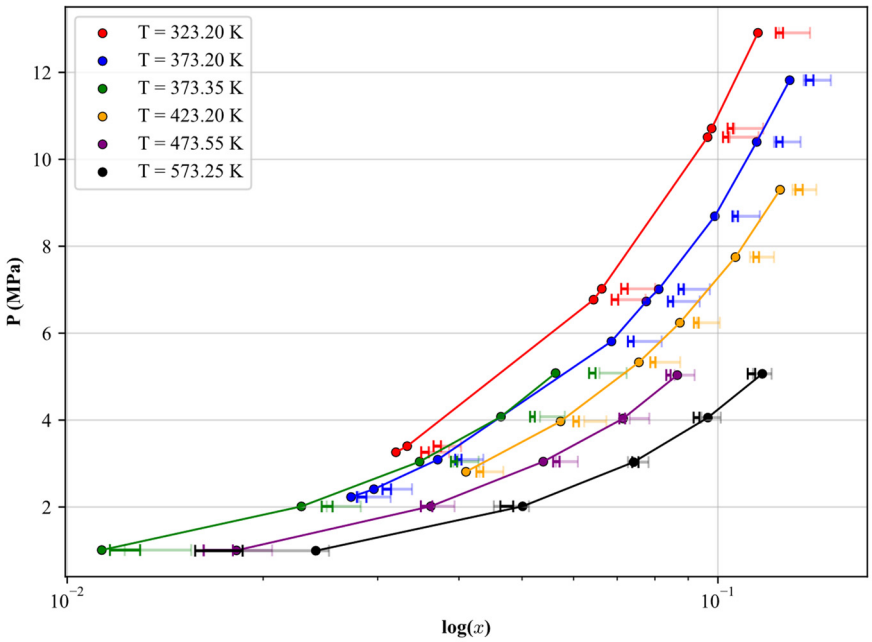

Figure 11. The 95% confidence interval for predictions of the n-eicosane samples.

In Figure 11, the experimental data points are represented by circles on the graph. The full-coloured intervals indicate the 95% confidence interval for model predictions in the first experiment, while the pale-coloured intervals correspond to the second experiment. A significant trend is observed, as the experimental data points align more closely with the full-coloured intervals, which suggests an improvement in model accuracy. Furthermore, the lengths of these full-coloured intervals are notably shorter than those of the pale-coloured intervals, reflecting an increased precision in prediction.

However, it is important to acknowledge that, despite the overall accuracy and precision, there are instances where the target values fall outside the prediction intervals. These occurrences underscore the limitations of the model and reveal areas where predictive errors persist. It is also important to note that the experiment was designed to demonstrate the effects of randomness, and therefore only a limited number of epochs were considered.

4. Summary and Conclusions

In conclusion, this study sought to leverage the capabilities of three distinct DNN models to predict H_2 solubility across a diverse range of n-alkanes. To achieve this, we gathered a comprehensive dataset that includes data for 15 different n-alkanes, sourced from publicly available resources. The key insights derived from our investigation can be summarised as follows:

- We employed a group-wise data partitioning approach to divide the dataset into training, validation, and testing sets, consisting of 9, 3, and 3 n-alkanes, respectively. Notably, the testing chemicals exhibited satisfactory performance, highlighting the adaptability of our developed models to novel n-alkanes not included in the current dataset.

- Our analysis of three model structures revealed that a DNN relying exclusively on dense layers is particularly prone to overfitting. Importantly, the integration of dropout layers effectively mitigated this issue.
- The DNN 3 model, characterised by its incorporation of batch normalisation and dropout layers, along with distinct input types, demonstrated remarkable performance. This was evidenced by an *RMSE* of 0.004 and an *SMAPE* of 2.58% on the validation dataset.
- The predictive performance of single DN models was notably comparable to that of ensemble methods, such as RF and GB, within the context of our study's database. This significant improvement can be attributed not only to the inherent characteristics of the models, which effectively mitigate overfitting, but also to our unique data partitioning strategy.
- The stability experiment conducted revealed that the implemented data-splitting scheme produces consistent predictions across multiple training trials. This finding underscores the robustness of our model's performance, even in the presence of potential variations during the training phase. Furthermore, the analysis of prediction confidence intervals demonstrated a remarkably high level of precision.
- Our study enhances the understanding of H_2 solubility across various chemical compositions, which is crucial for multiple industrial sectors, particularly in H_2-based renewable energy facilities. These advancements contribute to the safe and efficient design of H_2-based systems, promoting cleaner fuels and improving overall sustainability in energy production.

Our findings suggest promising avenues for further research, which could significantly contribute to sustainability efforts. We recommend investigating hyperparameter optimisation, exploring various learning rate decay scenarios, and considering transfer learning. Additionally, testing different characteristic properties, such as group contribution methods or chemical descriptors, is advisable. With the acquisition of additional experimental data in the future, there is potential to refine the weights of pre-trained models, thereby improving their precision and accuracy. These advancements will support the development of more efficient H_2-based systems, fostering cleaner energy technologies and promoting a transition to sustainable energy solutions.

Author Contributions: Conceptualization: A.T. and A.S.; Methodology: A.T. and A.S.; Investigation: A.T. and A.S.; Validation: A.T. and A.S.; Writing—Original Draft Preparation: A.T. and A.S.; Resources: A.Z. and M.H.; Supervision: A.Z. and M.H. All authors have read and agreed to the published version of the manuscript.

Funding: The authors did not receive support from any organisation for the submitted work.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data generated or analysed during this study are included in this article.

Conflicts of Interest: There are no conflicts of interest to declare in relation to this work.

Nomenclature

	Parameters
MSE	Mean squared error
MW	Molecular weight, g/mol
N	Number of data samples
P	Pressure, bar
P_C	Critical pressure, bar
P_D	Reduced or dimensionless pressure
Q_1	First quartile

Q_3	Third quartile
Res_i	Residual
$RMSE$	Root-mean-squared error
SD	Standard deviation
$SMAPE$	Symmetric mean absolute percentage error
SPE	Symmetric percentage Error
T	Temperature, K
T_B	Boiling point temperature, K
T_C	Critical temperature, K
T_D	Reduced or dimensionless temperature
t_i	Target value
V_C	Critical volume, mL/mol
x	H_2 solubility, mole fraction
y_i	Model output
Z_C	Critical compressibility factor
ρ_C	Critical density, g/mL
ω	Acentric factor
	Abbreviations
API	Application programming interface
EoS	Equation of states
DNN	Deep neural network
DT	Decision tree
ET	Extremely randomised trees
GB	Gradient boosting
H_2	Hydrogen gas
IQR	Interquartile range
ML	Machine learning
NN	Neural network
QQ	Quantile–Quantile
RF	Random forest
SVM	Support vector machine

Appendix A. Python Packages and System Information

Table A1 provides an overview of the Python packages utilised in this study, while Table A2 outlines the specifications of the computer system employed for this research.

Table A1. The main used Python packages with their corresponding release version.

Package	Version
python [54]	3.9.16
numpy [55]	1.26.4
pandas [56]	2.0.3
matplotlib [57]	3.7.2
seaborn [58]	0.12.2
sklearn [59]	1.3.0
scipy [60]	1.11.1
tensorflow [61]	2.10.1
cudnn [62]	8.2.1
cuda [63]	11.3
keras [64]	2.10.0

Table A2. Specifications of the used system for modelling.

Part	Specifications
CPU	11th Gen Intel(R) Core (TM) i7-11370H @ 3.30 GHz
RAM	31 GB
GPU	NVIDIA GeForce RTX 3050 Ti Laptop GPU
SSD	Total Size: 943 GB

References

1. Dawood, F.; Anda, M.; Shafiullah, G.M. Hydrogen production for energy: An overview. *Int. J. Hydrogen Energy* **2020**, *45*, 3847–3869. [CrossRef]
2. Pacheco, M.A.; Dassori, C.G. Hydrocracking: An improved Kinetic Model and Reactor Modeling. *Chem. Eng. Commun.* **2002**, *189*, 1684–1704. [CrossRef]
3. Florusse, L.J.; Peters, C.J.; Pàmies, J.C.; Vega, L.F.; Meijer, H. Solubility of hydrogen in heavy n-alkanes: Experiments and saft modeling. *AIChE J.* **2003**, *49*, 3260–3269. [CrossRef]
4. Lal, D.; Otto, F.D.; Mather, A.E. Solubility of hydrogen in Athabasca bitumen. *Fuel* **1999**, *78*, 1437–1441. [CrossRef]
5. Huang, S.H.; Lin, H.M.; Tsai, F.N.; Chao, K.C. Solubility of synthesis gases in heavy n-paraffins and Fischer-Tropsch wax. *Ind. Eng. Chem. Res.* **1988**, *27*, 162–169. [CrossRef]
6. Wilhelmsen, Ø.; Aasen, A.; Skaugen, G.; Aursand, P.; Austegard, A.; Aursand, E.; Gjennestad, M.A.; Lund, H.; Linga, G.; Hammer, M. Thermodynamic Modeling with Equations of State: Present Challenges with Established Methods. *Ind. Eng. Chem. Res.* **2017**, *56*, 3503–3515. [CrossRef]
7. von Solms, N.; Kouskoumvekaki, I.A.; Michelsen, M.L.; Kontogeorgis, G.M. Capabilities, limitations and challenges of a simplified PC-SAFT equation of state. *Fluid Phase Equilibria* **2006**, *241*, 344–353. [CrossRef]
8. Span, R.; Wagner, W.; Lemmon, E.W.; Jacobsen, R.T. Multiparameter equations of state—Recent trends and future challenges. *Fluid Phase Equilibria* **2001**, *183*, 1–20. [CrossRef]
9. Tatar, A.; Esmaeili-Jaghdan, Z.; Shokrollahi, A.; Zeinijahromi, A. Hydrogen solubility in n-alkanes: Data mining and modelling with machine learning approach. *Int. J. Hydrogen Energy* **2022**, *47*, 35999–36021. [CrossRef]
10. Zhang, R.; Chen, Y.; Li, Z.; Jiang, T.; Li, X. Two-stage robust operation of electricity-gas-heat integrated multi-energy microgrids considering heterogeneous uncertainties. *Appl. Energy* **2024**, *371*, 123690. [CrossRef]
11. Shang, Y.; Li, S. FedPT-V2G: Security enhanced federated transformer learning for real-time V2G dispatch with non-IID data. *Appl. Energy* **2024**, *358*, 122626. [CrossRef]
12. Zhu, N.; Wang, Y.; Yuan, K.; Yan, J.; Li, Y.; Zhang, K. GGNet: A novel graph structure for power forecasting in renewable power plants considering temporal lead-lag correlations. *Appl. Energy* **2024**, *364*, 123194. [CrossRef]
13. Esmaeili-Jaghdan, Z.; Tatar, A.; Shokrollahi, A.; Bon, J.; Zeinijahromi, A. Machine learning modelling of dew point pressure in gas condensate reservoirs: Application of decision tree-based models. *Neural Comput. Appl.* **2024**, *36*, 1973–1995. [CrossRef]
14. Foroughizadeh, P.; Shokrollahi, A.; Tatar, A.; Zeinijahromi, A. Hydrogen solubility in different chemicals: A modelling approach and review of literature data. *Eng. Appl. Artif. Intell.* **2024**, *136*, 108978. [CrossRef]
15. Benham, A.L.; Katz, D.L. Vapor-liquid equilibria for hydrogen–light-hydrocarbon systems at low temperatures. *AIChE J.* **1957**, *3*, 33–36. [CrossRef]
16. Sagara, H.; Arai, Y.; Saito, S. Vapor-Liquid Equilibria of Binary and Ternary Systems Containing Hydrogen and Light Hydrocarbons. *J. Chem. Eng. Jpn.* **1972**, *5*, 339–348. [CrossRef]
17. Tsang, C.Y.; Clancy, P.; Calado, J.C.G.; Streett, W.B. Phase Equilibria in The H2/CH4 System at Temperatures From 92.3 to 180.0 K and Pressures to 140 MPa. *Chem. Eng. Commun.* **1980**, *6*, 365–383. [CrossRef]
18. Hong, J.H.; Kobayashi, R. Vapor-liquid equilibrium study of the hydrogen-methane system at low temperatures and elevated pressures. *J. Chem. Eng. Data* **1981**, *26*, 127–131. [CrossRef]
19. Heintz, A.; Streett, W.B. Phase Equilibria in the H2/C2H4System at Temperatures from 114.1 to 247.1 K and Pressures to 600 MPa. *Berichte Bunsenges. Phys. Chem.* **1983**, *87*, 298–303. [CrossRef]
20. Burriss, W.L.; Hsu, N.T.; Reamer, H.H.; Sage, B.H. Phase Behavior of the Hydrogen-Propane System. *Ind. Eng. Chem.* **1953**, *45*, 210–213. [CrossRef]
21. Trust, D.B.; Kurata, F. Vapor-liquid phase behavior of the hydrogen-propane and hydrogen-carbon monoxide-propane systems. *AIChE J.* **1971**, *17*, 86–91. [CrossRef]
22. Nelson, E.E.; Bonnell, W.S. Solubility of Hydrogen in n-Butane. *Ind. Eng. Chem.* **1943**, *35*, 204–206. [CrossRef]
23. Aroyan, H.J.; Katz, D.L. Low Temperature Vapor-Liquid Equilibria in Hydrogen-n-Butane System. *Ind. Eng. Chem.* **1951**, *43*, 185–189. [CrossRef]
24. Klink, A.E.; Cheh, H.Y.; Amick Jr, E.H. The vapor-liquid equilibrium of the hydrogen—N-butane system at elevated pressures. *AIChE J.* **1975**, *21*, 1142–1148. [CrossRef]
25. Freitag, N.P.; Robinson, D.B. Equilibrium phase properties of the hydrogen—Methane—Carbon dioxide, hydrogen—Carbon dioxide—N-pentane and hydrogen—N-pentane systems. *Fluid Phase Equilibria* **1986**, *31*, 183–201. [CrossRef]

26. Connolly, J.F.; Kandalic, G.A. Gas solubilities, vapor-liquid equilibria, and partial molal volumes in some hydrogen-hydrocarbon systems. *J. Chem. Eng. Data* **1986**, *31*, 396–406. [CrossRef]
27. Brunner, E. Solubility of hydrogen in 10 organic solvents at 298.15, 323.15, and 373.15 K. *J. Chem. Eng. Data* **1985**, *30*, 269–273. [CrossRef]
28. Gao, W.; Robinson, R.L.; Gasem, K.A.M. Solubilities of Hydrogen in Hexane and of Carbon Monoxide in Cyclohexane at Temperatures from 344.3 to 410.9 K and Pressures to 15 MPa. *J. Chem. Eng. Data* **2001**, *46*, 609–612. [CrossRef]
29. Lachowicz, S.K.; Newitt, D.M.; Weale, K.E. The solubility of hydrogen and deuterium in n-heptane and n-octane at high pressures. *Trans. Faraday Soc.* **1955**, *51*, 1198–1205. [CrossRef]
30. Cook, M.W.; Hanson, D.N.; Alder, B.J. Solubility of Hydrogen and Deuterium in Nonpolar Solvents. *J. Chem. Phys.* **1957**, *26*, 748–751. [CrossRef]
31. Peramanu, S.; Pruden, B.B. Solubility study for the purification of hydrogen from high pressure hydrocracker off-gas by an absorption-stripping process. *Can. J. Chem. Eng.* **1997**, *75*, 535–543. [CrossRef]
32. Connolly, J.F.; Kandalic, G.A. Thermodynamic properties of solutions of hydrogen in n-octane. *J. Chem. Thermodyn.* **1989**, *21*, 851–858. [CrossRef]
33. Kim, K.J.; Way, T.R.; Feldman, K.T.; Razani, A. Solubility of Hydrogen in Octane, 1-Octanol, and Squalane. *J. Chem. Eng. Data* **1997**, *42*, 214–215. [CrossRef]
34. Prausnitz, J.M.; Benson, P.R. Solubility of liquids in compressed hydrogen, nitrogen, and carbon dioxide. *AIChE J.* **1959**, *5*, 161–164. [CrossRef]
35. Sebastian, H.M.; Simnick, J.J.; Lin, H.-M.; Chao, K.-C. Gas-liquid equilibrium in the hydrogen + n-decane system at elevated temperatures and pressures. *J. Chem. Eng. Data* **1980**, *25*, 68–70. [CrossRef]
36. Schofield, B.A.; Ring, Z.E.; Missen, R.W. Solubility of hydrogen in a white oil. *Can. J. Chem. Eng.* **1992**, *70*, 822–824. [CrossRef]
37. Park, J.; Robinson, R.L., Jr.; Gasem, K.A.M. Solubilities of Hydrogen in Heavy Normal Paraffins at Temperatures from 323.2 to 423.2 K and Pressures to 17.4 MPa. *J. Chem. Eng. Data* **1995**, *40*, 241–244. [CrossRef]
38. Gao, W.; Robinson, R.L.; Gasem, K.A.M. High-Pressure Solubilities of Hydrogen, Nitrogen, and Carbon Monoxide in Dodecane from 344 to 410 K at Pressures to 13.2 MPa. *J. Chem. Eng. Data* **1999**, *44*, 130–132. [CrossRef]
39. Lin, H.-M.; Sebastian, H.M.; Chao, K.-C. Gas-liquid equilibrium in hydrogen + n-hexadecane and methane + n-hexadecane at elevated temperatures and pressures. *J. Chem. Eng. Data* **1980**, *25*, 252–254. [CrossRef]
40. Breman, B.B.; Beenackers, A.A.C.M.; Rietjens, E.W.J.; Stege, R.J.H. Gas-Liquid Solubilities of Carbon Monoxide, Carbon Dioxide, Hydrogen, Water, 1-Alcohols (1. ltoreq. n. ltoreq. 6), and n-Paraffins (2. ltoreq. n. ltoreq. 6) in Hexadecane, Octacosane, 1-Hexadecanol, Phenanthrene, and Tetraethylene Glycol at Pressures up to 5.5 MPa and Temperatures from 293 to 553 K. *J. Chem. Eng. Data* **1994**, *39*, 647–666.
41. Luo, H.; Ling, K.; Zhang, W.; Wang, Y.; Shen, J. A Model of Solubility of Hydrogen in Hydrocarbons and Coal Liquid. *Energy Sources Part A Recovery Util. Environ. Eff.* **2010**, *33*, 38–48. [CrossRef]
42. Abdi, J.; Bastani, D.; Abdi, J.; Mahmoodi, N.M.; Shokrollahi, A.; Mohammadi, A.H. Assessment of competitive dye removal using a reliable method. *J. Environ. Chem. Eng.* **2014**, *2*, 1672–1683. [CrossRef]
43. Tatar, A.; Shokrollahi, A.; Halali, M.A.; Azari, V.; Safari, H. A Hybrid Intelligent Computational Scheme for Determination of Refractive Index of Crude Oil Using SARA Fraction Analysis. *Can. J. Chem. Eng.* **2015**, *93*, 1547–1555. [CrossRef]
44. Setzmann, U.; Wagner, W. A New Equation of State and Tables of Thermodynamic Properties for Methane Covering the Range from the Melting Line to 625 K at Pressures up to 100 MPa. *J. Phys. Chem. Ref. Data* **1991**, *20*, 1061–1155. [CrossRef]
45. Douslin, D.R.; Harrison, R.H. Pressure, volume, temperature relations of ethane. *J. Chem. Thermodyn.* **1973**, *5*, 491–512. [CrossRef]
46. Thomas, R.H.P.; Harrison, R.H. Pressure-volume-temperature relations of propane. *J. Chem. Eng. Data* **1982**, *27*, 1–11. [CrossRef]
47. Ambrose, D.; Tsonopoulos, C. Vapor-Liquid Critical Properties of Elements and Compounds. 2. Normal Alkanes. *J. Chem. Eng. Data* **1995**, *40*, 531–546. [CrossRef]
48. Brunner, E.; Hültenschmidt, W.; Schlichthärle, G. Fluid mixtures at high pressures IV. Isothermal phase equilibria in binary mixtures consisting of (methanol + hydrogen or nitrogen or methane or carbon monoxide or carbon dioxide). *J. Chem. Thermodyn.* **1987**, *19*, 273–291. [CrossRef]
49. Lemmon, E.W.; Goodwin, A.R.H. Critical Properties and Vapor Pressure Equation for Alkanes C_nH2_{n+2}: Normal Alkanes with $n \leq 36$ and Isomers for n = 4 through n = 9. *J. Phys. Chem. Ref. Data* **2000**, *29*, 1–39. [CrossRef]
50. Yaws, C.L. Chapter 1—Critical Properties and Acentric Factor—Organic Compounds. In *Thermophysical Properties of Chemicals and Hydrocarbons*, 2nd ed.; Yaws, C.L., Ed.; Gulf Publishing Company: Oxford, UK, 2014; pp. 1–124.
51. Shokrollahi, A.; Safari, H.; Esmaeili-Jaghdan, Z.; Ghazanfari, M.H.; Mohammadi, A.H. Rigorous modeling of permeability impairment due to inorganic scale deposition in porous media. *J. Pet. Sci. Eng.* **2015**, *130*, 26–36. [CrossRef]
52. Ioffe, S.; Szegedy, C. In Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Mountain View, CA, USA, 6 July 2015; pp. 448–456.
53. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
54. Rossum, V. *Python Tutorial*; Technical Report CS-R9526; Centre for Mathematics and Computer Science: Amsterdam, The Netherlands, 1995.

55. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array programming with NumPy. *Nature* **2020**, *585*, 357–362. [CrossRef]
56. McKinney, W. Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010; pp. 51–56.
57. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [CrossRef]
58. Waskom, M.; Botvinnik, O.; O'Kane, D.; Hobson, P.; Lukauskas, S.; Gemperline, D.C.; Augspurger, T.; Halchenko, Y.; Cole, J.B.; Warmenhoven, J. *Mwaskom/Seaborn: V0. 8.1*; Zenodo: Boston, MA, USA, 2017.
59. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
60. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef] [PubMed]
61. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv* **2016**, arXiv:1603.04467.
62. Chetlur, S.; Woolley, C.; Vandermersch, P.; Cohen, J.; Tran, J.; Catanzaro, B.; Shelhamer, E. cudnn: Efficient primitives for deep learning. *arXiv* **2014**, arXiv:1410.0759.
63. NVIDIA, P.V.; Fitzek, F.H. Cuda, Release: 11.3. 2023. Available online: https://developer.nvidia.com/cuda-toolkit (accessed on 1 July 2024).
64. Chollet, F. Keras. 2015. Available online: https://github.com/fchollet/keras (accessed on 1 July 2024).