*sustainability*

MDPI

*Article*

# Machine Learning-Based Dry Gas Reservoirs Z-Factor Prediction for Sustainable Energy Transitions to Net Zero

**Progress Bougha [1], Foad Faraji [1,*], Parisa Khalili Nejad [2], Niloufar Zarei [3], Perk Lin Chong [1], Sajid Abdullah [1], Pengyan Guo [4] and Lip Kean Moey [5]**

[1] School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough TS1 3BX, UK; progressbougha83@gmail.com (P.B.); p.chong@tees.ac.uk (P.L.C.); s.abdullah@tees.ac.uk (S.A.)
[2] Kellogg College, Oxford University, Oxford OX2 6PN, UK; parisa.khalilinejad@kellogg.ox.ac.uk
[3] VaasaETT, Fredrikinkatu 47, 00100 Helsinki, Finland; niloufar.zarei@vaasaett.com
[4] School of Mechanical Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450045, China; guopengyan@ncwu.edu.cn
[5] Centre for Modelling and Simulation, Faculty of Engineering, Built Environment & Information Technology, SEGi University, Petaling Jaya 47810, Malaysia; moeylipkean@segi.edu.my
* Correspondence: f.faraji@tees.ac.uk

## Abstract

Dry gas reservoirs play a pivotal transitional role in meeting the net-zero target worldwide. Accurate modelling and simulation of this energy source require fast and reliable prediction of the gas compressibility factor (Z-factor). The experimental measurements of Z-factor are the most reliable source; however, they are expensive and time-consuming. This makes developing accurate predictive models essential. Traditional methods, such as empirical correlations and Equations of States (EoSs), often lack accuracy and computational efficiency. This study aims to address these limitations by leveraging the predictive power of machine learning (ML) techniques. Hence in this study three ML models of Artificial Neural Network (ANN), Group Method of Data Handling (GMDH), and Genetic Programming (GP) were developed. These models were trained on a comprehensive dataset comprising 1079 samples where pseudo-reduced pressure (Ppr) and pseudo-reduced temperature (Tpr) served as input and experimentally measured Z-factors as output. The performance of the developed ML models was benchmarked against two cubic EoSs of Peng–Robinson (PR) and van der Waals (vdW), and two semi-empirical correlations of Dranchuk-Abou-Kassem (DAK) and Hall and Yarborough (HY), and recent developed ML based models, using statistical metrics of Mean Squared Error (MSE), coefficient of determination ($R^2$), and Average Absolute Relative Deviation Percentage (AARD%). The proposed ANN model reduces average prediction error by approximately 70% relative to the PR equation of state and by over 35% compared with the DAK correlation, while maintaining robust performance across the full Ppr and Tpr of dry gas systems. Additionally paired *t*-tests and Wilcoxon signed-rank tests performed on the ML results confirmed that the ANN model achieved statistically significant improvements over the other models. Moreover, two physical equations using the white-box models of GMDH and GP were proposed as a function of Ppr and Tpr for prediction of the dry gas Z-factor. The sensitivity analysis of the data shows that the Ppr has the highest positive effect of 88% on Z-factor while Tpr has a moderate effect of 12%. This study presents the first unified, statistically validated comparison of ANN, GMDH, and GP models for accurate and interpretable Z-factor prediction. The developed models can be used as an alternative tool to bridge the limitation of cubic EoSs and limited accuracy and applicability of empirical models.



Academic Editor: Hamid Arastoopour

Received: 8 December 2025
Revised: 12 January 2026
Accepted: 26 January 2026
Published: 8 February 2026

**Keywords:** dry gas; Z-factor; machine learning; net zero transition; equation of state (EoS)

## 1. Introduction

Dry gas mostly consists of methane with negligible amounts of liquid, typically emits 50–60% less $CO_2$ per unit of energy than coal and 30% less than oil. Hence, they play an important transitional role in sustainable development and the effort to meet global net-zero goals. As shown in Figure 1, dry natural gas production has been increasing since 2015 and has alone reached about 143 billion cubic feet worldwide in 2023, underscoring its importance in energy provision [1]. Hence, the development and optimisation of these reservoirs are critical owing to their high market value and significant contribution to the oil and gas industry. Accurate prediction of the gas compressibility factor is central to pressure–volume–temperature (PVT) analysis and underpins volumetric calculations, separator design, and flow modelling in gas production and transmission systems [2–4]. To provide a quantitative operational context, a pipeline case study reported by [5] is particularly informative. In this case study, a dry-gas stream was transported through a 30-mile, 6-inch pipeline at 1200 psi inlet pressure and production of 30 MMscf/day under isothermal conditions, and the outlet pressure was computed using a steady-state gas-flow equation in which Z enters linearly. Keeping pipeline geometry, flow rate, and thermodynamic conditions fixed, the authors varied only the Z-factor model and quantified the resulting outlet-pressure deviation. They reported outlet-pressure errors of 59 psi (4 bar) for Beggs–Brill and 37 psi (2.5 bar) for Hall–Yarborough relative to Standing–Katz chart, whereas their ML-based Z-factor model reduced the deviation to 10 psi (<1 bar). These results demonstrate that improving Z-factor accuracy can materially reduce uncertainty in pressure-drop calculations that inform compressor duty and energy-use estimates, which is essential for sustainable practices and reduce the emission. The full calculation procedure of the case study can be found in [5].

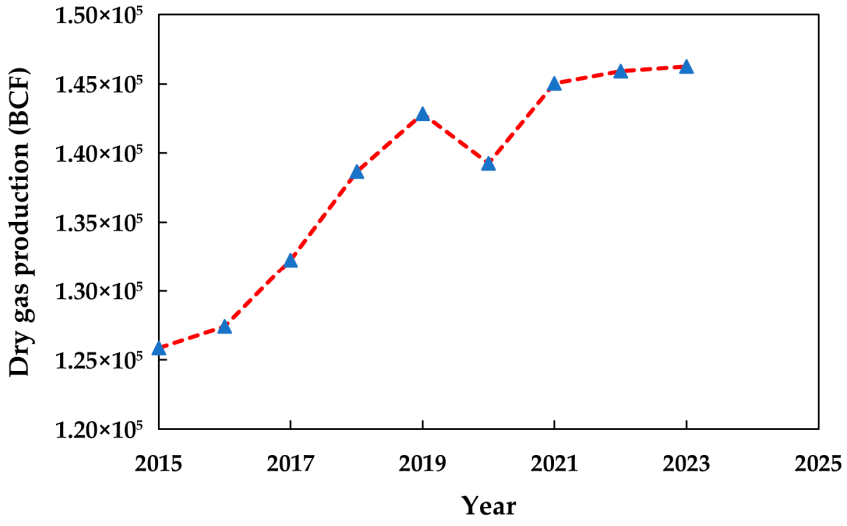



**Figure 1.** Dry natural gas production worldwide [1].

Hence accurate determination of the Z-factor is crucial for reliable calculation and characterisation of dry gas reservoirs. The Z-factor represents the ratio between the actual volume of a gas at a given pressure and temperature and the volume it would occupy under ideal-gas conditions [6]. This is mathematically expressed as

$$Z = \frac{PV}{nRT} \quad (1)$$

where $P$ is the pressure, $V$ is the volume, n is the number of moles, $R$ is the gas constant, and $T$ is the temperature [7]. If the assumption that all perfect gases follow the same PVT behaviour at the same dimensionless Pr and Tr values is correct, then all real gases with the same reduced pressure and temperature should have similar Z-factor [8]. The gas Z-factor is approximately 1 under standard pressure and temperature conditions. At low pressures and high temperatures, the Z-factor remains close to unity, indicating that the gas behaves nearly ideally [9].

Despite its importance, obtaining accurate Z-factor values remains challenging. The constant volume depletion (CVD) laboratory test is the method most commonly used for Z-factor determination [10]. Experimental measurements expect to produce the accurate results, but expensive and time-consuming, as determining the Z-factor for each gas composition at every pressure and temperature requires extensive laboratory testing [11]. Furthermore, on many occasions, sufficient samples of fluids cannot be obtained, particularly at an early production stage, which makes it challenging to use the experimental approach. Equations of State (EoS) are widely used to predict gas compressibility and other thermodynamic properties. However, their application can be cumbersome due to the need for numerous parameters, such as binary interaction coefficients, enthalpy, entropy, Gibbs free energy, and acentric factors, which are not always available for all components [12,13]. Among the various proposed EoSs, the cubic forms of van der Waals (1873) [14], Soave–Redlich–Kwong (SRK) [15] and Peng–Robinson (PR)1976 [3] remain the most common for natural-gas systems. The PR model, which accounts for both molecular attraction and repulsion and includes a temperature-dependent correction term, is thermodynamically consistent and generally reliable for dry-gas reservoirs [13]. The general form of cubic EoS for the prediction of natural gas Z-factor is as follows.

$$Z^3 + a \times Z^2 + b \times Z + c = 0 \quad (2)$$

where $a$, $b$, and $c$ are empirical coefficients, and $Z$ denotes to the compressibility factor. The conventional calculation procedure of the PR EoS is shown in Appendix A.

Because EoS formulations can be computationally demanding and sometimes inaccurate outside their calibration range, several empirical and semi-empirical correlations have been developed to estimate the gas deviation factor (Z). The Standing–Katz (1942) chart [16] laid the foundation for numerous analytical expressions derived from experimental data [9,17–23]. Among these, the Hall and Yarborough (1973) [17] and Dranchuk–Abou-Kassem (1975) [24] correlations are the most widely adopted, offering improved accuracy across extended pressure and temperature ranges.

Despite these advances, both EoS and empirical correlations exhibit limitations when applied to diverse dry-gas compositions or extreme reservoir conditions [25]. Their predictive accuracy depends strongly on tuned coefficients and restricted datasets. This gap motivates the present study, which applies data-driven machine-learning (ML) models to develop a more general and accurate prediction framework for the Z-factor of dry gas reservoirs.

In recent years, ML techniques have become one of the most powerful tools for modelling complex relationships across various industries [26–33]. On the natural dry gas Z-factor, several publications have effectively proved the ability of different ML algorithms to improve the accuracy and reliability of predictions against traditional empirical correlations and EoSs. Ref. [34] employed Genetic Programming (GP) to develop a novel correlation for predicting the natural and sour gas mixtures Z-factor. Their model was based on a substantial experimental dataset of 977 gas samples, over wide ranges of Ppr and Tpr. Ref. [35] extracted 6638 data points from the Standing and Katz Z-factor diagram and developed several ANNs as a function of Ppr and Tpr for the prediction of Z-factor. Then, Ref. [36] utilised 5500 data points for modelling natural gas Z-factor using Ppr and Tpr as

inputs and proved the accuracy of the ANN over EoSs and empirical correlations. Then in another study Ref. [37] developed natural gas Z-factor models based on Least Square Support Vector Machine (LSSVM) by utilising 4753 data sets. Using 978 data points [4] applied Wilcoxon generalised radial basis function neural network (WGRBFN) for prediction of natural gas Z-factor as a function of Ppr, Tpr. Ref. [38] utilised ANN for the prediction of natural gas Z-factor using SK chart data. Ref. [11] argued that although the accuracy of the above ML developed models was very good, they act like a black-box and no meaningful correlation can be derived from them. Hence, they used 978 data points to develop a group method of data handling (GMDH) neural network and developed a mathematical correlation for estimating natural gas Z-factor. Yet, they introduced a GMDH comparisons with broader modelling strategies remain limited. The existing research in this area is fragmented and typically focuses on single models or specific hybrid frameworks (e.g., ANN–GA or evolutionary–ML combinations) without a unified, head-to-head evaluation across different ML paradigms [29,37,39]. Previous studies have not jointly assessed ANN, GMDH, and GP on the same dry-gas databank, nor have they systematically benchmarked these models against both cubic equations of state (EoSs) and widely used empirical correlations. The present study addresses these gaps by deploying GMDH-type neural networks, GP, and a conventional ANN for dry-gas Z-factor estimation, and by performing a comprehensive comparative analysis against existing EoSs and empirical correlations, while prioritising model transparency and symbolic output to enhance applicability in operational and design contexts. Hence, the contributions of this study are as follows: (i) a unified, head-to-head evaluation of ANN, GMDH, and GP on a single 1079-point dry-gas databank; (ii) statistical rigour via 10-fold cross-validation and formal significance tests (paired *t*-tests and Wilcoxon signed-rank tests); (iii) interpretable, white-box equations from GMDH/GP rather than black-box ML alone; and (iv) deployment-readiness, with models designed for direct embedding in reservoir simulators and PVT workflows.

## 2. Materials and Methods

### *2.1. Databank*

The reliability and accuracy of any data-driven model are directly influenced by the quality of the databank used for training and testing [40,41]. Accordingly, a comprehensive databank comprising 1079 data points was compiled from open literature sources [42–46]. Of the total dataset, 80% was used for model training and the remaining 20% for testing model performance. The dataset contains two independent variables of Ppr and Tpr as well as derived Z-factors, obtained from the constant volume depletion (CVD) laboratory experiment, for each value of Ppr and Tpr. To ensure consistency and reproducibility, all pressures were converted to mega pascal and all temperatures to kelvin. Pseudo-critical properties (Pc, Tc) were estimated using Kay's mixing rules as Ppr = P/Pc and Tpr = T/Tc. Z-factors were then obtained by interpolation from the Standing–Katz correlation using the corresponding (Ppr, Tpr) pairs.

Data quality checks included unit-consistency verification and statistical screening using a z-score threshold of $\pm 3$ evaluated within Tpr bands to avoid masking temperature-dependent trends. The analysis indicated that no outliers were detected, and all samples were retained. The dataset spans $0.2 \leq \text{Ppr} \leq 30$ and $1.05 \leq \text{Tpr} \leq 3$ and is provided in Supplementary Materials (https://doi.org/10.5281/zenodo.18225906) for reproducibility.

The databank provides a comprehensive understanding of the Z-factor of various dry gas mixtures under a varying set of conditions. The summary of the data bank, its sources and validity range are presented in Table 1, while the detailed statistical description is provided in Table 2. As it can be seen the dataset encompasses a comprehensive range of values for the measured Ppr, Tpr, and Z values, allowing for tests under various conditions.

**Table 1.** Summary of literature sources and operating conditions for the 1079-point Z-factor databank used in this study.

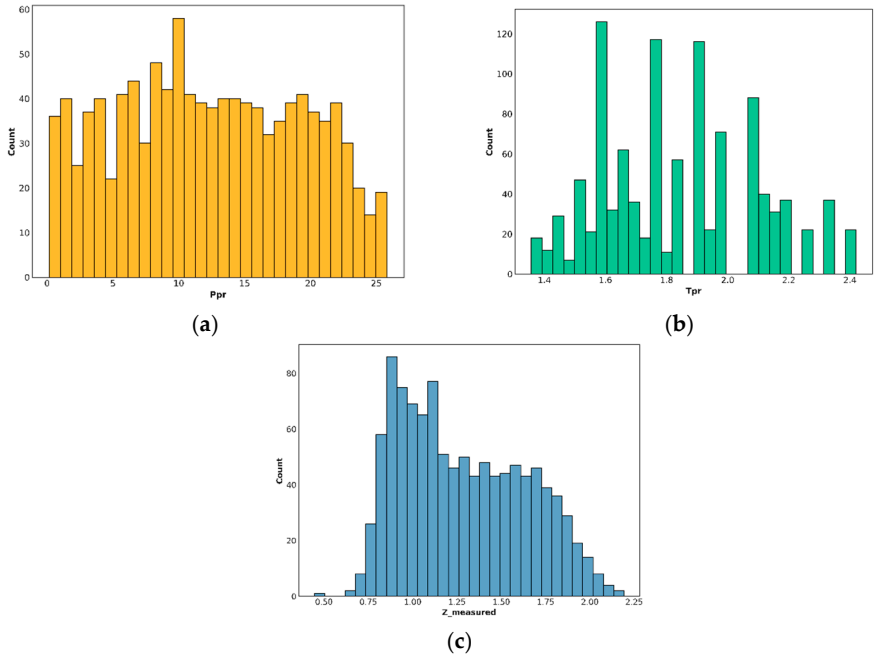
| Reference | Ppr Range | Tpr Range | Z Measured | Mixture |
|---|---|---|---|---|
| [42] | 1.38–10.21 | 1.35–1.80 | 0.445–1.14 | Lean Natural gas |
| [41] | 2.19–25.33 | 1.60–2.42 | 0.92–2.03 | Gas-condensate at dry state |
| [44] | 0.16–1.59 | 1.42–1.90 | 0.86–0.99 | Natural dry gas |
| [45] | 4.78–25.82 | 1.60–2.35 | 0.87–2.19 | Gas-condensate at dry state |
| [46] | 1.34–9.98 | 1.39–1.72 | 0.71–1.14 | Dry gas |

**Table 2.** The statistical comparison the utilised data in this study.

| Statistic | Ppr | Tpr | Z Measured |
|---|---|---|---|
| Count | 1079 | 1079 | 1079 |
| Mean | 12.361 | 1.838 | 1.285 |
| Std [1] | 6.90 | 0.25 | 0.351 |
| Minimum | 0.162 | 1.357 | 0.445 |
| 25% | 6.799 | 1.627 | 0.979 |
| 50% | 12.059 | 1.828 | 1.230 |
| 75% | 18.239 | 2.074 | 1.575 |
| Maximum | 25.821 | 2.420 | 2.192 |

[1] Standard Deviation.

Histograms of Tpr, Ppr, and the Z-factor are shown in Figure 2. The Ppr histogram exhibits a relatively uniform spread with a slight concentration around 10–15, indicating well-distributed pressure data. The Tpr histogram is more skewed, with values clustering between 1.6 and 2.0, suggesting a narrower range of reduced temperatures. In contrast, the histogram of measured Z shows a moderately right-skewed distribution, with most values concentrated between 0.9 and 1.3, reflecting compressibility behaviour under typical reservoir conditions. Overall, Ppr spans a wide range with a relatively uniform distribution from 0.162 to 25.82, indicating a dataset that covers both low- and high-pressure scenarios.




(**a**) (**b**) (**c**)

**Figure 2.** The histogram shown the distribution of the data set utilised in this study (**a**) pseudo reduced pressure (Ppr); (**b**) Pseudo-reduced temperature (Tpr) and (**c**) calculated Z-factor from compositional data.

The relationships, strengths, and directions among the variables were analysed and are presented as a heatmap in Figure 3. The results show a very strong positive correlation between Ppr and the Z-factor (r = 0.96), while Tpr exhibits a moderate positive correlation with the measured Z-factor (r = 0.34). These findings support the validity of using Ppr and Tpr as inputs for predicting the Z-factor.

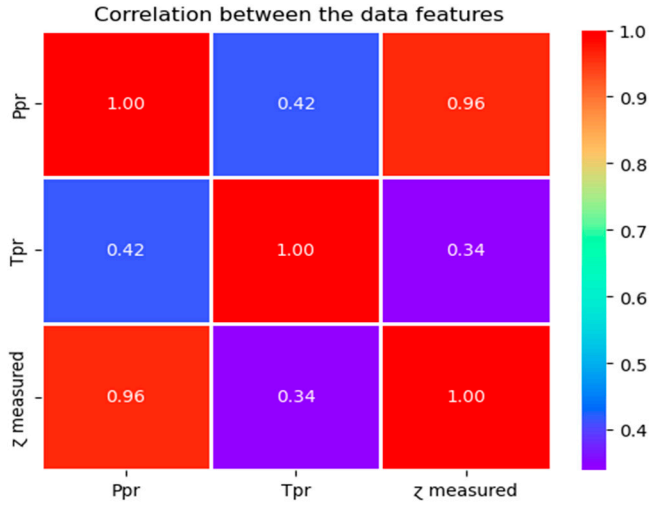



**Figure 3.** Correlation analysis of the variables in the databank.

### *2.2. Artificial Neural Network (ANN)*

Inspired by the structure of the human central nervous system, artificial neural networks (ANNs) are powerful generalisation algorithms that have gained widespread use across many fields. The general architecture of the ANN is shown in Figure 4, which consists of three inputs, hidden and output layers. In the input layer, each node represents a feature, $x_1, \ldots, x_n$, which are passed forward to the hidden layer without any computation. Then in the hidden layer, each input is multiplied by a corresponding weight, $w_1, \ldots, w_n$, and summed together along with a bias unit. Then, the weighted sum is processed by typically non-linear activation function such as sigmoid or rectified linear unit (ReLU) [47]. The output from the activation function is then passed to the output layer, where it becomes the final network output. Once the network is established, then it will be trained to adjust the weights and the bias unit based on the error between the predicted values and the expected values, using the backpropagation or gradient decent approach [48]. A network with only one hidden layer can be represented mathematically as follows.

$$y = f\left(\Sigma_{i=1}^{n} w_i x_i + b\right) \quad (3)$$

where y is the output of the neuron, $f$ is the activation function, $w_i$ are the weights, $x_i$ are the input features and b is the bias term.

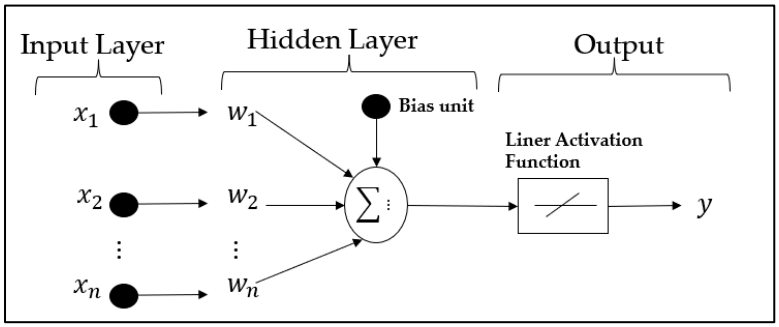



**Figure 4.** The general architecture of feedforward artificial neural network.

In this study, a typical multilayer artificial neural network (ANN) architecture shown in Figure 5, was implemented. The best topology in terms of the number of hidden layers and neurons per layer is problem-dependent and not uniquely defined [49]. However, a single hidden-layer feedforward network has been shown to provide sufficient approximation capability for many engineering applications [32]. Accordingly, single hidden-layer networks with the number of neurons varying from 5 to 70 (in increments of 5) were evaluated using a grid-search approach. The ANN input layer consisted of two neurons corresponding to pseudo-reduced pressure (Ppr) and pseudo-reduced temperature (Tpr), while the output layer contained a single neuron representing the gas Z-factor. A ReLU activation function was used in the hidden layer, and a linear activation function was adopted for the output layer. Three well-known backpropagation training algorithms of Levenberg–Marquardt (LM), Bayesian Regularisation (BR), and Scaled Conjugate Gradient (SCG), were examined during model development [50]. Based on cross-validated performance, the ANN trained using the LM algorithm with 45 hidden neurons (2–45–1 architecture) was selected as the final configuration, as it provided the lowest prediction error and highest coefficient of determination.

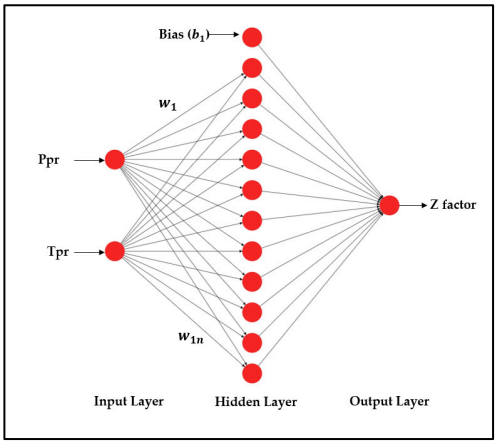



**Figure 5.** Architecture of the feedforward ANN developed for predicting the dry gas Z-factor.

Mathematically, ReLU is defined as follows:

$$ReLU(x) = \max(0, x) \tag{4}$$

The function operates by returning $x$ when $x > 0$ and zero when $x \leq 0$. In this case $x$ could be any real number, which coming from the weight sum of the outputs from the previous layer of neurons plus a bias term.

To avoid overfitting and ensure the robustness of the developed ANNs, a 10-fold cross-validation framework was incorporated into the modelling procedure. The databank (N = 1079) was normalised using min–max scaling prior to model training. Specifically, Ppr and Tpr inputs and the Z-factor output were scaled using min–max normalisation over the full dataset. A grid search over hidden-layer size (5–70 neurons in steps of 5) was evaluated using 10-fold cross-validation (K = 10) across the normalised dataset. For each neuron count, a feedforward ANN trained with LM, BR and SCG was fitted on nine folds and evaluated on the remaining fold; performance metrics were averaged over the ten folds to select the optimal hidden-layer size.

After selecting the optimal neuron count, a separate hold-out split was performed (HoldOut = 0.2) with a fixed random seed (rng = 1), and the selected ANN configuration was retrained on the 80% training subset and evaluated on the 20% testing subset. Each ANN training run used a single random weight initialisation.

Additionally, sensitivity analysis was performed to quantify the relative influence of Ppr and Tpr on Z-factor prediction using Garson's algorithm, in which, for each hidden neuron i, the contribution of input j is computed as

$$C_{i,j} = |W1(i,j)| \times |W2(i)| \quad (5)$$

where W1 and W2 are the weight matrices connecting the input layer to the hidden layer and the hidden layer to the output layer, respectively. The total contribution of input j is then obtained by summing over all hidden neurons as follows:

$$I_j = \sum_{i=1}^{H} C_{i,j} \quad (6)$$

The relative importance of each input is then calculated as follows:

$$RI_j = 100 \times \frac{I_j}{\sum_{k=1}^{n} I_k} \quad (7)$$

A summary of the final ANN architecture and hyperparameters is provided in Table 3.

**Table 3.** The selected hyperparameters of the final ANN model applied in this study.

| Hyperparameter | Value |
|---|---|
| Number of inputs | Ppr, Tpr |
| Hidden layers | 1 |
| Hidden neurons | 45 |
| Output | Z-factor |
| Hidden activation function | ReLU |
| Output activation | Linear |
| Training algorithm | LM |
| Neuron search range | 5–70 (step 5) |
| Initialisations | 1 |
| Validation | 10-fold CV + 80/20 hold-out |

### *2.3. Group Method of Data Handling (GMDH)*

In this study, the GMDH algorithm was implemented to develop a robust empirical correlation for predicting the Z-factor of natural gas mixtures using Ppr and Tpr. The GMDH is a self-organising neural network that constructs a predictive model through successive layers of polynomial combinations [51]. In the first stage of the GMDH procedure, all possible pairs of input variables were generated, and for each pair, a second-order polynomial model was constructed in the following form:

$$Z_i^{GMDH} = ax_i + bx_j + ax_ix_j + dx_i^2 + ex_j^2 + fx_ix_j \quad (8)$$

This is the classical Kolmogorov–Gabor polynomial equation where $Z$ represents the dry gas Z-factor, $x_i$ and $x_j$ are the inputs parameters paired together (Ppr and Tpr), $a$, $b$, $e$, and $f$ are coefficients of the equation. The outputs of these models, denoted by $z_i$, form a new matrix in the form of $v_z = (Z_1, Z_2, \ldots Z_n)$, which serves as new input for next layer. Then the coefficients of each polynomial model are determined using the least squares method as follows.

$$\delta^2 = \sum_{i=1}^{Nt} \left(y_i - Z_i^{GMDH}\right)^2 \quad j = 1, 2 \ldots, \binom{M}{2} \quad (9)$$

where $Nt$ is the number of the data points in the training set and $M$ is number of input variables. To compute the unknown polynomial coefficients efficiently, the general model is written in matrix form as follows:

$$A^T = YX^T\left(XX^T\right)^{-1} \tag{10}$$

where $X$ is the matrix of input features, $Y$ is the vector of target values, and $A$ contains the model coefficients. Once the coefficients were estimated from the training data, the selected pairs were evaluated on the testing data as follows.

$$\delta^2 = \sum_{i=Nt+1}^{Nt}\left(y_i - Z_i{}^{GMDH}\right)^2 < \varepsilon \quad j = 1,\ 2\ldots,\ \binom{M}{2} \tag{11}$$

Only those models that satisfy the error threshold $\varepsilon$ are propagated to the next layer. The algorithm proceeds iteratively, generating new layers of models from the best-performing candidates until the prediction error no longer improves or begins to increase, at which the iteration stops [52]. Unlike standard neural network models, the GMDH approach generates a transparent, physically interpretable equation suitable for direct mathematical implementation. To ensure that GMDH model is statistically robust, sensitivity analysis was performed, and model configurations were evaluated to determine the optimal structure. Polynomial orders from 2 to 4 and neuron limits from 20 to 80 per layer were tested. Performance was assessed using $R^2$, MSE, and AARD%. The quadratic GMDH structure with a maximum of 50 neurons per layer consistently produced the best results while preventing overfitting. Therefore, this configuration was adopted for the final model. A summary of the final GMDH architecture and hyperparameters is provided in Table 4.

**Table 4.** The final hyperparameter of the GMDH model.

| Hyperparameter | Value |
|---|---|
| Number of inputs | Ppr, Tpr |
| Output | 0Z-factor |
| Polynomial order (tested) | 2, 3, 4 |
| Selected polynomial order | $p$ = X (replace this) |
| Max neurons per layer (tested) | 20, 40, 50, 60, 80 |
| Selected max neurons per layer | $N_{max}$ = (replace this) |
| Maximum number of layers | 5 |
| Error Threshhold ($\varepsilon$) | 0.005 |
| Neuron Selection criteria | MSE |
| Training method | Least-squares (pseudo-inverse) |
| Normalisation | Min–max scaling |
| Data split | 80% development/20% validation |
| Random seed | 1 |

### *2.4. Genetic Programming*

Genetic programming, originally introduced by Koza [53], is an evolutionary algorithm derived from the principles of natural selection and biological evolution. It is designed to evolve computer programmes that solve user-defined tasks by iteratively improving a population of candidate solutions. Unlike conventional neural networks, which often function as black-box models with limited interpretability, GP stands out for

its ability to produce transparent, symbolic solutions, making it especially useful for modelling nonlinear and complex problem. In GP, solutions are typically represented as tree structures, as illustrated in Figure 6. Each tree encodes a candidate mathematical expression or program [54]. The process begins with the random initialisation of a population of trees. These trees have then evolved over successive generations through biologically inspired operations such as selection, crossover, and mutation. Selection chooses the fittest individuals, crossover generates offspring by combining parent trees, and mutation introduces random variation to maintain population diversity [39]. A key advantage of GP over traditional nonlinear regression is its ability to simultaneously discover both the structure and the parameters of the model. Whereas nonlinear regression requires a predefined functional form that is incrementally adjusted during optimisation, GP dynamically generates and evolves the form and the content of the model without prior assumptions [55]. The core objective is to evolve a symbolic function, composed of arithmetic operators, variables, and mathematical transformations, that maps a set of input features to an output. This relationship is generally expressed as

$$y = f\ (x_1,\ x_2,\ \ldots,\ x_n) \tag{12}$$

where y is the general output of the model that GP try to find, which is a function of input variables of $x_1, x_2, \ldots, x_n$. An evolutionary optimisation process with mutation, crossover, and selection yields an explicit analytical expression that best fits the experimental data, as given below.

$$Z = \frac{aPpr + bTpr + c}{d + ePpr^2 + fTpr^2} \tag{13}$$

where $a, b, c, d, e, f$ are coefficient of the equation, determined through training. The GP model parameters were selected based on a sensitivity analysis in which the population size and number of generations were varied between 500 and 3000. The model performance improved rapidly up to approximately 1500–2000 individuals and then reached a level, with no significant gains beyond this range. To balance accuracy and computational cost, the final GP configuration used 2000 trees, a maximum tree depth of 6, and a mutation–crossover of 0.1. The details of the GP parameters developed in this study are shown in Table 5.

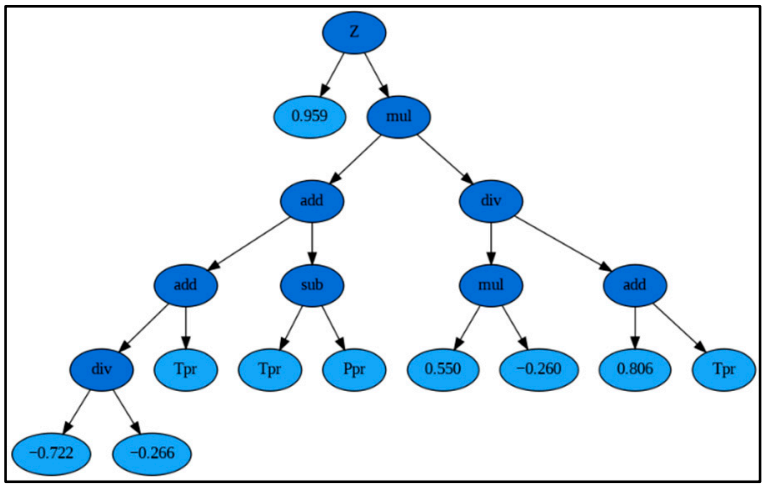



**Figure 6.** The developed genetic programming model for the prediction of dry gas Z-factor. mul: multiplication; div: division; sub: subtraction.

**Table 5.** The detailed parameter selection of the GP for prediction of dry gas Z-factor.

| Parameter | Value |
|---|---|
| Population size (No trees) | 2000 |
| Generation | 50 |
| Stopping criteria | 0.01 |
| p_crossover [1] | 0.7 |
| p_subtree mutation | 0.1 |
| p_hoist mutation | 0.05 |
| p_point mutation | 0.1 |
| Maximum samples | 0.9 |

[1] p stands for probability.

### *2.5. Model Performance Evaluation*

The accuracy of the developed models and utilised existing literature models was evaluated using three statistical parameters of Mean Square Error (*MSE*), coefficient of determination ($R^2$) and Average Absolute Relative Deviation (*AARD%)* shown below.

$$MSE = \frac{1}{n}\Sigma_{i=1}^{n}\left(Z_{iexp} - Z_{ipred}\right)^2 \quad (14)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Z_{iexp} - Z_{ipred}\right)^2}{\sum_{i=1}^{n}\left(Z_{iexp} - Z_{im-exp}\right)^2} \quad (15)$$

$$AARD\% = \frac{1}{n}\Sigma_{i=1}^{n}\left|\frac{Z_{iexp} - Z_{ipred}}{Z_{iexp}}\right| \times 100 \quad (16)$$

In the above equations, $n$ is the number of data points, and the subscripts *exp*, *m-exp*, and pred represent the experimental Z-factor, the mean of the experimental Z-factor, and the model-predicted Z-factor values, respectively. The testing data sets is used to evaluate the performance of the intelligent models. It is important for the models to perform well in predicting unseen data sets, hence the testing data sets used to evaluate the performance of the models. In addition to statistical error analysis the graphical methods in the form of cross-plot and bar graph were used to assess the performance of the developed models. Moreover, a trend analysis was conducted to assess whether the models correctly follow the physical behaviour of the experimental Z-factor across different pressure ranges.

## 3. Results and Discussion

In this study, three intelligent models of ANN, GMDH, and GP were developed to predict the dry gas Z-factor as a function of Ppr and Tpr. For this purpose, a comprehensive databank of 1079 data points from the literature, covering a wide range of Ppr, Tpr, and Z-factor values, was employed. The data were obtained from compositional analyses of various dry-gas reservoirs composed predominantly of methane with trace amounts of non-hydrocarbon gases. The conventional Kay's mixing rule was applied to calculate Ppr and Tpr, and the Standing–Katz (SK) chart was used to estimate the Z-factor. During data preparation, 80% of the dataset was allocated for training and the remaining 20% for testing the developed models. To determine the best ANN topology, a single hidden-layer network with an arbitrary number of neurons between 5 and 70 was examined. A systematic grid-search procedure was employed to identify the optimal ANN architecture. Hidden-layer sizes from 5 to 70 neurons (increments of 5) were evaluated using 10-fold cross-validation, and the best configuration was selected based on the lowest MSE and highest $R^2$ [32,49]. Furthermore, three optimisation algorithms of LM, BR and SCG were employed to tune the weights and biases of the network. Figure 7 presents the results of the ANN models

trained with various optimisation techniques and different neuron counts in the single hidden layer.

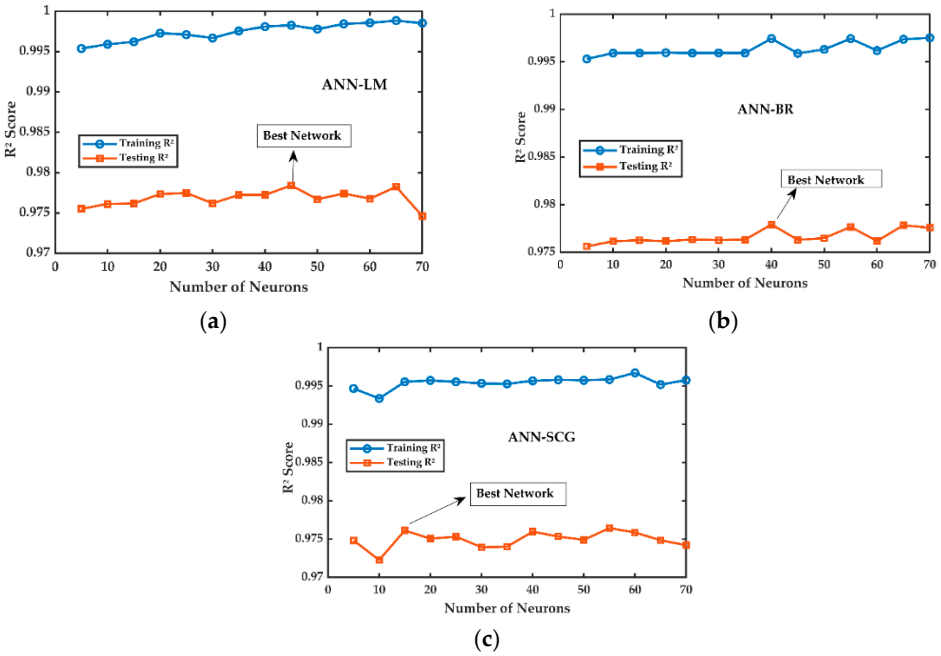



**Figure 7.** Performance of the single-hidden-layer ANN models trained using (**a**) Levenberg–Marquardt (LM), (**b**) Bayesian regularisation (BR), and (**c**) scaled conjugate gradient (SCG) for natural-gas Z-factor prediction. A grid search over hidden-layer size (5–70 neurons, step = 5) was conducted for each training algorithm, and the optimal topology was selected based on the highest cross-validated coefficient of determination ($R^2$). Key takeaway: ANN-LM with 45 neurons achieved the best overall fit and was therefore used as the primary ANN configuration in subsequent analyses.

All developed ANN architectures exhibited high predictive performance across both the training and testing datasets. The difference between the testing and the training errors is almost negligible, demonstrating that the network avoids the inherent overfitting problem [56]. After comparing all developed ANNs, the ANN optimised with the LM algorithm (ANN-LM) with 45 neurons in the hidden layer (2–45–1) was identified as the best model with lowest MSE of 0.002808 and AARD% of 1.68, and the highest $R^2$ of 0.9770. Therefore, this model was selected for predicting the dry gas Z-factor alongside the other methods. To avoid overfitting and ensure the robustness of the developed ANN, a 10-fold cross-validation framework was implemented. In addition to the ANN model, two further intelligent approaches of GMDH and GP were employed, yielding the explicit correlations presented in Equations (17) and (18), respectively. The performance of the developed models was evaluated against three cubic EoSs of vdW, SRK, and PR; two empirical correlations (HY and DAK) and three previously ML based models using MSE, AARD%, and $R^2$, with the results presented in Table 4.

$$\begin{aligned} Z_{dry\ gas} = {} & 0.2343 - 1.1803Ppr + 0.3646Tpr + 6.3681Ppr^2 - 0.3563Tpr^2 \\ & - 0.6015PprTpr - 6.6270Ppr^3 + 0.0345Tpr^3 \\ & - 1.9582Ppr^2Tpr + 1.3677PprTpr^2 + 2.3352Ppr^4 \\ & + 0.0621Tpr^4 + 1.6193Ppr^3Tpr - 0.2540PprTpr^3 \\ & - 0.5903Ppr^2Tpr^2 \end{aligned} \quad (17)$$

$$Z_{dry\ gas} = \left( \frac{\left[ \frac{(Tpr - Ppr)}{(-1.775)} - (Ppr + 2.075) \right]}{\left[ \frac{(5.752 Tpr)}{(Ppr + 2.075)} - (Tpr + 5.155 + Tpr/3.646) \right]} \right) \quad (18)$$

When compared with the literature models, the proposed ANN–LM demonstrates superior predictive accuracy ($R^2$ = 0.9987; MSE = 0.0028) on a larger and more diverse dataset. The results confirm that the present unified framework yields higher generalisation capability than previously published MLFN and MLP-based ANN models while maintaining statistical rigour through 10-fold cross-validation and formal significance testing. It is worth mentioning that although DAK outperformed the two intelligent models, GMDH and GP, in terms of AARD%, it was still less accurate with respect to MSE and $R^2$. The HY correlation not only produced relatively large errors in predicting the experimental Z-factor but was also applicable to only 776 points from our databank, being restricted to the range 1.1 < Tpr < 3 and 0.2 < Ppr < 15. By contrast, the DAK correlation covers a wider range of 0.7 < Tpr < 3 and 0.01 < Ppr < 30, which explains its better performance and its ability to predict the entire dataset. The statistical error metrics were computed using the subset of data points for which each classical correlation is applicable within its recommended validity range (e.g., the Hall–Yarborough correlation was evaluated on 776 data points).

Figure 8 illustrates the cross-plots of the best-developed ANN optimised with LM, as well as the GMDH and GP models, for predicting the dry gas Z-factor in this study. As can be seen from Figure 8a, the ANN-LM model exhibits excellent performance in both the training and testing phases for natural gas Z-factor prediction. The majority of the data points lie close to the perfect-fit line, with only two test points falling outside this region. The cross-plot of the GMDH model (Figure 8b) also shows a good fit for both the training and testing sets, with some scatter around the fit line at lower and higher Z-factor values. A similar trend is observed for the GP model (Figure 7c), where scatter around the perfect-fit line is more evident in the low and high Z-factor regions.

In addition to the statistical performance metrics presented in Table 6, we conducted formal statistical significance tests to determine whether the improvements achieved by the ANN model over the other methods is meaningful across the full range of the dataset. Specifically, paired *t*-tests and Wilcoxon signed-rank tests were applied to compare the absolute prediction errors between ANN and other models. The statistical analysis demonstrated that the ANN model exhibits significantly different error behaviour compared with all classical and empirical models evaluated. Paired *t*-tests and Wilcoxon signed-rank tests confirmed that the ANN's prediction errors were significantly lower than PR with $p \approx 10^{-151}$, vdW with $p \approx 10^{-108}$, DAK with $p \approx 10^{-61}$, and GP with $p \approx 10^{-4}$, indicating that the probability of these improvements occurring by chance is effectively zero. When compared with the Hall–Yarborough correlation within its validity range, ANN also showed markedly smaller errors with $p \approx 10^{-139}$, verifying better performance even in the region where HY performs best. The only model statistically indistinguishable from ANN was GMDH with $p > 0.3$, reflecting the close similarity in accuracy between the two intelligent techniques. Region-based significance tests further showed that ANN consistently outperformed PR across low, moderate, and high Ppr ranges with $p < 10^{-7}$ to $p < 10^{-69}$. Although paired *t*-tests and Wilcoxon signed-rank tests indicate statistically significant differences between models ($p < 0.05$), the practical significance of these differences depends on their magnitude. While the ANN-LM model achieves the lowest overall error, the predictive performance of the ANN-LM and other recent ANN-based models reported in the literature differs by less than 1% in AARD, indicating very close practical accuracy.

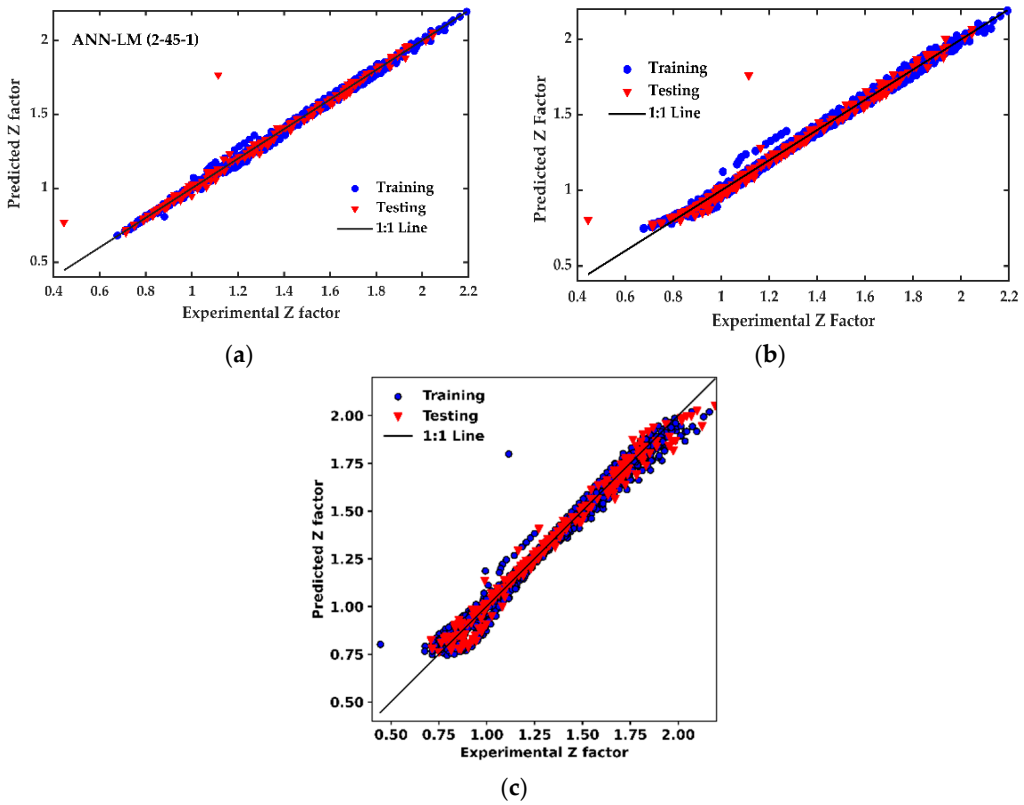



**Figure 8.** Predicted versus experimental Z-factor cross-plots for the developed dry-gas models: (**a**) ANN–LM, (**b**) GMDH, and (**c**) GP. Blue circles and red triangles represent the training (80%) and testing (20%) subsets, respectively; the solid black line indicates 1:1 agreement. Key takeaway: ANN–LM exhibits the tightest clustering around the 1:1 line (lowest scatter) across the full Z range, whereas GMDH and GP show wider dispersion, particularly in the mid-range Z values.

**Table 6.** The statistical comparison of the developed and utilised literature models for prediction of dry gas Z-factor.

| Model | AARD% | MSE | $R^2$ | Dataset (No) | Note |
|---|---|---|---|---|---|
| ANN-LM (2–45–1) | 1.68 | 0.002766 | 0.9987 | 1079 | Proposed unified ANN (LM); statistically validated |
| GMDH | 4.60 | 0.00637 | 0.9480 | 1079 | Symbolic white-box model |
| GP | 5.87 | 0.0949 | 0.9365 | 1079 | Symbolic regression (GP) |
| DAK | 2.65 | 0.2633 | 0.8748 | 1036 | Conventional correlation |
| HY [1] | 29.63 | 0.74042 | 0.0393 | 1000 | Conventional correlation |
| van der Walls | 9.47 | 0.03029 | 0.9837 | --- | Classical cubic EOS |
| PR | 5.57 | 0.01477 | 0.9864 | --- | Classical cubic EOS |
| MLFN (ANN) [57] | 1.98 | --- | 0.979 | 1079 | Dry-gas dataset, same inputs |
| GMDH [11] | 2.88 | 0.00115 (From RMSE) | 0.9176 | 978 | GMDH model (no statistically significant variation) |
| MLP-ANN [58] | --- | 0.014544 (from RMSE) | 0.9903 | 604 | Black box models (MLP-ANN and RB-ANN), (no statistically significant variation |

[1] Only valid for 776 data points.

To further investigate the performance of the developed models in various pressures, the graph of the relative error against the Ppr were plotted and presented in Figure 9. Among all, the ANN-LM model (Figure 9a) demonstrates the highest accuracy, maintaining minimal and tightly clustered relative errors across all pressure region in the data bank. The genetic programming model (Figure 9c) shows moderate performance with

some scatter, particularly Ppr < 10, while GMDH (Figure 9b) exhibits larger deviations, especially at low and mid-range Ppr < 15, indicating reduced reliability. The DAK model (Figure 9d) performs well where majority of the data is within +/−10% error, yet there are underprediction within the 10 < Ppr < 15. In contrast, HY correlations underpredicted the Z-factor in lower pressure region of 0 < Ppr < 15. The van der Waals EoS (Figure 9g) exhibits the poorest performance, with a consistent underprediction across the entire pressure range. Nevertheless, the Peng–Robinson cubic EoS predicts the dry gas Z-factor with good accuracy at low pressures but increasingly overpredicts as the pressure rises. These results underscore the superior predictive capability of machine learning–based model, particularly ANN-LM, compared with empirical correlations and conventional EoSs in the whole data range. Compared with previously published ANN and GMDH-based models [11,57,58], the proposed ANN–LM (2–45–1) achieved the lowest AARD (1.68%) and highest $R^2$ (0.9987) across a larger and more diverse dry-gas dataset (1079 points), confirming superior accuracy, robustness, and statistical validity over the existing literature models. Nevertheless, the GMDH model reported in the literature [11] exhibits slightly better performance than the GMDH model developed in this study, which may be attributed to the smaller and more homogeneous dataset (978 points) used in that work, leading to reduced variability but limited generalisation.

The performance of the best developed model in predicting the dry gas Z-factor was tested using an independent data bank collected from [35], which was outside the data that was used in this study. The performance of the ANN-LM in predicting the Z-factor as a function of Ppr is tested and presented in Figure 10. The result shows that the developed ANN-LM model can follow the physical trend of experimental data accuracy, indicating a strong performance of the mode. This confirms that the model follows the expected physical behaviour, as increasing the pressure will increase the dry gas Z-factor.

To verify that the symbolic equation of GMDH and GP were not overfitted and maintained physically plausible behaviour, a comprehensive robustness and sensitivity analysis was performed. For the GMDH model, polynomial orders from 2 to 4 and neuron limits between 20 and 80 per layer were evaluated. The optimal quadratic structure with a maximum of 50 neurons per layer achieved stable convergence, with comparable training and testing $R^2$ values of 0.948 vs. 0.944), respectively, indicating the absence of overfitting. Similarly, for the GP model, population size and number of generations were varied between 500 and 3000, and the model performance plateaued beyond approximately 2000 individuals, confirming robustness of the evolved symbolic expressions. Coefficient sensitivity was further tested by perturbing the normalised input variables (Ppr and Tpr) within ±10%, which resulted in a maximum variation of 3–5% in the predicted Z-factor, demonstrating low parameter sensitivity and numerical stability.

The physical plausibility of both symbolic models was examined graphically across the full Ppr range (Figure 9), where predictions correctly follow the expected real-gas behaviour, approaching Z = 1 as Ppr = 0 and increasing monotonically with Ppr at fixed Tpr. At high-pressure limits (Ppr ≈ 30), both GMDH and GP models remained smooth and free from error fluctuation, confirming that polynomial overfitting was effectively mitigated and that the developed symbolic equations capture physically consistent trends across the entire operating domain.

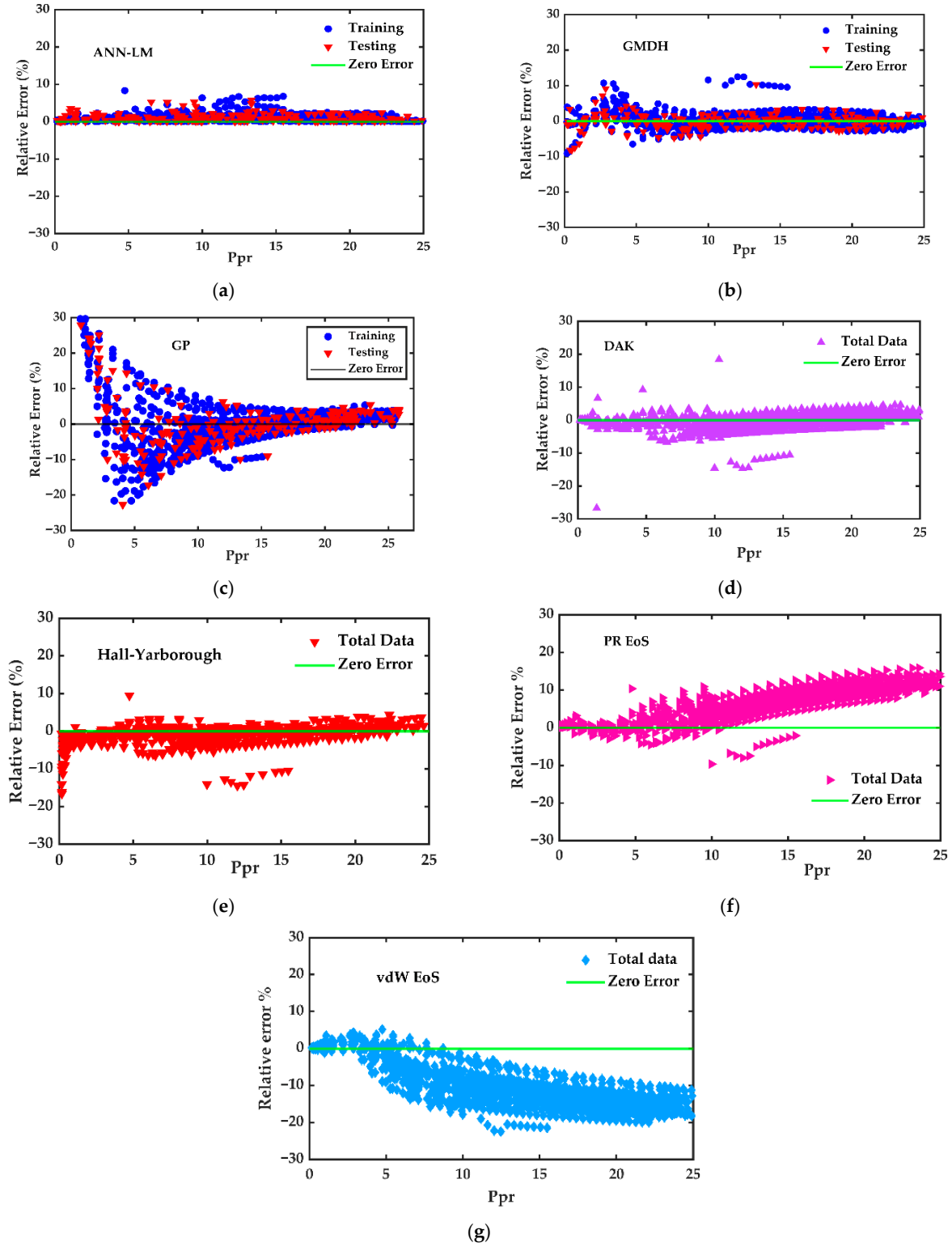



**Figure 9.** Relative error (%) versus pseudo-reduced pressure (Ppr) for the developed models and benchmark correlations: (**a**) ANN–LM, (**b**) GMDH, (**c**) GP, (**d**) Dranchuk–Abou-Kassem (DAK), (**e**) Hall–Yarborough (HY), (**f**) Peng–Robinson EOS (PR), and (**g**) van der Waals EOS (vdW). For ANN–LM, GMDH, and GP, blue circles and red triangles denote the training (80%) and testing (20%) subsets; for DAK, HY, PR, and vdW, symbols represent the full dataset. The green line indicates zero error. Key takeaway: ANN–LM maintains the smallest and most stable error band across the full Ppr range, whereas the empirical correlations and EOS show systematic pressure-dependent bias, with errors increasing at higher Ppr.

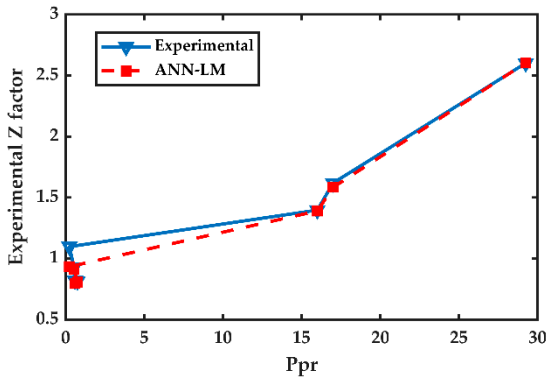



**Figure 10.** The trend analysis of ANN-LM prediction against the experimental values as a function of Ppr.

The Garson-based sensitivity analysis shown in Figure 11, indicates that Ppr accounts for approximately 88% of the total input importance, while Tpr contributes about 12%. This confirms that pseudo-reduced pressure is the dominant factor controlling the Z-factor in the present dataset. This finding is consistent with gas compressibility behaviour, where pressure exerts a stronger and more direct influence on deviation from ideal-gas conditions, particularly in the high-pressure region. The smaller contribution of Tpr reflects the comparatively lower variability of temperature in the dataset and its secondary effect on real-gas behaviour. Overall, these results demonstrate that the ANN correctly identifies Ppr as the primary driver of Z-factor changes for dry gases under the studied conditions. From a practical standpoint, the proposed ANN, GMDH and GP models can be directly embedded into reservoir simulation and field workflows as fast surrogates for conventional EoS-based Z-factor calculations. The explicit polynomial forms of the GMDH and GP models can be implemented in existing PVT or simulator modules, while the trained ANN can be exported as a stand-alone function or lookup routine for use in production forecasting, well-test interpretation and real-time decision support where rapid and reliable Z-factor estimation is required.

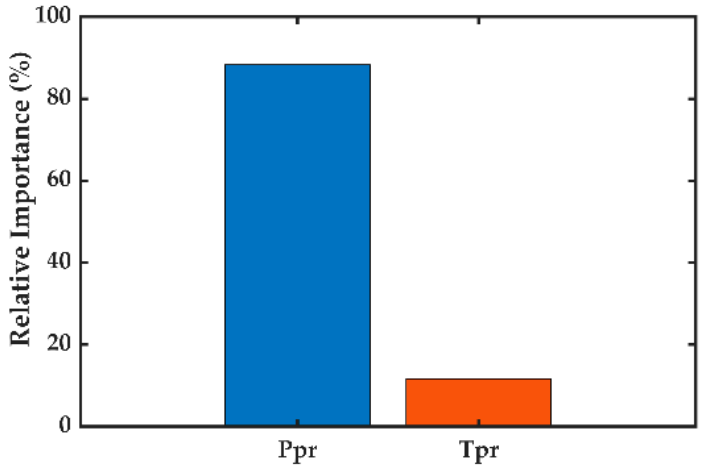



**Figure 11.** The impact of the input variables on the dry gas Z-factor using the Garson method.

The accuracy and computational efficiency of Z-factor models have direct implications for sustainable energy transitions. In natural-gas production and transmission, small errors in Z-factor estimation propagate into flowrate, pressure-drop, and energy-consumption calculations [25]. The proposed ANN, GMDH, and GP models reduce prediction uncertainty by over 40% compared with conventional correlations, enabling more reliable resource management and process optimisation. Because the developed models evaluate the Z-factor

in a single forward pass without iterative EOS solving, computational cost per evaluation decreases by one to two orders of magnitude [5] facilitating real-time digital-twin simulations and energy-efficient control. When implemented in compressor-station design or pipeline operation, this reduction in uncertainty can translate to measurable decreases in recompression energy demand and $CO_2$ emissions. Thus, accurate and interpretable Z-factor modelling supports cleaner, more efficient gas-processing systems and contributes to Sustainable Development Goals 7 and 13.

*Applicability and Limitations*

The developed models were trained on 1079 samples representing dry, methane-dominant natural gas systems. As shown in Supplementary Materials (https://doi.org/10.5281/zenodo.18225906 (data repository)), methane is the dominant component in all mixtures, while non-hydrocarbon gases are present only within limited ranges. $CO_2$ contents vary from trace levels up to approximately 20 mol%, hydrogen sulphide ($H_2S$) up to 9 mol%, and nitrogen is present only in minor fractions typical of dry-gas reservoirs.

In such systems, the influence of composition on gas compressibility is largely captured through pseudo-critical properties and the corresponding pseudo-reduced variables (Ppr and Tpr) [8]. This modelling choice is consistent with established Z-factor practice, where classical correlations (e.g., Standing–Katz-type approaches and related explicit correlations) represent real-gas deviation primarily as a function of reduced pressure and temperature. Accordingly, Ppr and Tpr provide a compact and physically meaningful input space for data-driven Z-factor modelling of dry and lean gas mixtures.

Consequently, the developed models are applicable primarily to lean and moderately sour dry gases, where gas behaviour can be reliably characterised using pseudo-reduced pressure and temperature. Because the models do not explicitly include compositional descriptors, predictive accuracy outside the trained compositional envelope, such as for $CO_2$-rich, $H_2S$-rich, or heavy-hydrocarbon-rich gases, cannot be guaranteed. Predictions should also be restricted to the trained thermodynamic domain ($0.2 \leq Ppr \leq 30$ and $1.05 \leq Tpr \leq 3.0$). Future extensions of the framework could incorporate compositional variables (or improved pseudo-critical corrections) to broaden applicability to richer or highly sour gas systems.

## 4. Conclusions

Accurate dry gas Z-factor estimation is a practical way for lowering the operational carbon intensity of gas systems. In this study, a comprehensive dataset of 1079 dry-gas samples from different reservoirs was used to develop and benchmark three machine-learning models of ANN, GMDH, and GP for prediction of the gas deviation factor as a function of pseudo-reduced pressure and temperature. Unlike previous works this study provides a unified and statistically rigorous assessment of multiple ML paradigms against classical cubic EoSs and empirical correlations. The main findings are summarised as follows:

1. The DAK and Hall–Yarborough correlations are constrained by their validity ranges and, over the full dataset, yield AARD% values between 2.64 and 29.63. While DAK performs better than HY, our analysis shows that both correlations are vulnerable to loss of accuracy when applied outside their original development envelopes.
2. Among the cubic equations of state evaluated, Peng–Robinson (PR) gave the best performance (AARD% of 5.57; MSE of 0.01477; $R^2$ of 0.9864), outperforming van der Waals (vdW) (AARD% of 9.47; MSE of 0.03029; $R^2$ of 0.9837). However, neither cubic EoS reproduced the Z-factor with sufficient accuracy across the full Ppr–Tpr domain, underscoring the need for more flexible surrogate models.

3. The ANN model optimised with the LM algorithm provides the highest overall accuracy with AARD% of 1.68, MSE of 0.0028 and $R^2$ of 0.9987 and successfully captures the expected physical trends of Z as a function of Ppr and Tpr. Statistical significance tests including paired t-test and Wilcoxon confirmed that ANN's error reduction relative to PR, vdW, DAK, HY, and GP is highly significant with $p < 0.001$, demonstrating that the improvement is not due to random variation.
4. The GMDH and GP models deliver explicit analytical expressions for Z-factor as a function of Ppr, Tpr, offering a novel, interpretable alternative to black-box ML approaches. These symbolic equations attain acceptable accuracy but exhibit some degradation in the low-pressure region (Ppr < 10), indicating a clear direction for future refinement.
5. Garson-based sensitivity analysis indicates that Ppr contributes 88% and Tpr 12% to the variation in the dry-gas Z-factor.
6. Overall, the novelty of this study lies in (i) the comprehensive and statistically validated comparison of multiple ML models with conventional Z-factor correlations and cubic EoSs over an extended dry-gas data bank, and (ii) the development of simulation-ready, symbolic GMDH and GP equations that can be directly embedded into reservoir simulators and PVT modules, offering reservoir engineers both accuracy and interpretability for practical field workflows
7. The models developed in this study are valid within the ranges 0.162 < Ppr < 25.821 and 1.357 < Tpr < 2.420; their accuracy is expected to decrease when applied beyond these limits.

**Supplementary Materials:** All the relevant supplementary materials can be found in https://doi.org/10.5281/zenodo.18225906 (accessed on 7 December 2025).

**Author Contributions:** Conceptualization, P.B. and F.F.; methodology, F.F. and P.B.; software, P.B. and F.F.; validation, F.F. and P.L.C., formal analysis, P.B. and F.F.; investigation, P.B. and F.F.; data curation, P.B., P.K.N. and F.F.; writing—original draft preparation, P.B. and F.F.; writing—review and editing, F.F., N.Z., P.G., L.K.M. and S.A.; visualisation, F.F., P.K.N. and N.Z.; supervision, F.F. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study were derived from the following resources available in the public domain: Compressibility Factors for Lean Natural Gas-Carbon Dioxide Mixtures at High Pressure [10.2118/1590-PA] [42]. Measurement and Calculation of Gas Compressibility Factor for Condensate Gas and Natural Gas under Pressure up to 116 MP [10.1016/j.jct.2013.03.02] [41]. A Study on the Supercompressibility and Compressibility Factors of Natural Gas Mixtures [10.1016/0920-4105(91)90016-G] [44]. Experiments and Modeling of Volumetric Properties and Phase Behavior for Condensate Gas under Ultra-High-Pressure Conditions [10.1021/ie2025757] [45]. Satter, A.; Campbell, J.M. Non-Ideal Behavior of Gases and Their Mixtures [10.2118/566-PA] [46]. The original data presented in the study are openly available in Natural Gas compressibility factor and associated ML-based codes (ANN+GMDH+GP)] at Supplementary Materials (https://doi.org/10.5281/zenodo.18225906).

**Conflicts of Interest:** Niloufar Zarei is an employee of Vaasaett (Finland). This employment did not influence the study design, data collection, analysis, interpretation, or decision to publish. The authors declare no conflicts of interest.

## Abbreviations

| | |
|---|---|
| ANN | Artificial Neural Network |
| LM | Levenberg–Marquardt |
| BR | Bayesian Regularisation |
| SCG | Scaled Conjugate Gradient |
| GMDH | Group method of data handling |
| GP | Genetic Programming |
| Ppr | Pseudo-reduced pressure |
| Tpr | Pseudo-reduced temperature |
| EoS | Equation of State |
| Tr | Critical temperature |
| Pr | critical pressure |
| R | universal gas constant |
| PVT | Pressure-Volume-Temperature |
| MMscf/day | Million Standard Cubic Feet per Day |

## Appendix A

The PR equation takes into account intermolecular forces in the form of attraction and repulsion, and the volume of the molecules. A temperature-dependent factor is introduced in the equation's term accounting for attractive forces between molecules, which in general form can be represented as follows.

$$P = \frac{RT}{V_m - b} - \frac{a(T)}{V_m(V_m + b) + b(V_m - b)} \tag{A1}$$

where $P$ is the pressure of the gas, $T$ is the temperature, $Vm$ is molar volume of the gas,

R is the universal gas constant, a and b are attraction and repulsive parameters, respectively. The temperature dependency correction factor, $a(T)$ can be determined as follows:

$$a(T) = a_c\left[1 + \left(0.37464 + 1.54226\omega - 0.26992\omega^2\right)\left(1 - \sqrt{T_r}\right)\right]^2 \tag{A2}$$

where $\omega$ is the acentric factor of the substance, and $Tr$ is reduced temperature. In this study we assumed the acentric factor of 0.011 for the dry gas compositions. The $a_c$ and $b$ in the above equations are dimensionless variables that can be determined as follows.

$$\begin{aligned} a_c &= 0.45724\frac{R^2{T_c}^2}{P_c} \\ b &= 0.07780\frac{RT_c}{P_c} \end{aligned} \tag{A3}$$

where $R$ is the universal gas constant, $Tc$ and $Pc$ are critical temperature and pressure, respectively. The PR equation is thermodynamically consistent, and thus it gives reliable predictions for phase behaviour, which makes it suitable for dry gas reservoirs [59]. The expression in Equation (3) can be rewrite for computing the Z-factor.

$$Z^3 - (1 - B)Z^2 + \left(A - 3B^2 - 2B\right)Z - \left(AB - B^2 - B^3\right) = 0 \tag{A4}$$

where $A$, $B$, are determined as follows:

$$A = \frac{a(T)P}{R^2{T_c}^2} \tag{A5}$$

$$B = \frac{bP}{RT} \tag{A6}$$

where a is attraction parameter at the critical point, b is repulsion parameter.

Hall and Yarborough (1973) [17] equation is based on the Carnahan–Starling equation-of-stat and uses a complex iterative procedure to improve the limits of the SK chart for the regions of Ppr > 15 and Tpr < 1 [13].

$$Z = \frac{1 + y + y^2 - y^3}{(1-y)^3}\left(14.76t - 9.76t^2 + 4.58t^3\right) + \left(90.7t - 242.2t^2 + 42.4t^3\right) y^{1.1} \quad \text{(A7)}$$

where $t = Tpc/T$ and y is the reduced density. Then later in 1973, Dranchuk et al. (1973) [24] used Benedict–Webb–Rubin (Simon & Briggs, 1964) EoS [60] and proposed a Z-factor correlation. Their equation is modified by Abu-Kassem and Dranchuk (1975) [24] who added a reduced density parameter and developed an implicit analytical equation for estimation of the Z-factor. This method, known as DAK equation, uses polynomial expressions to relate reduced pressure, temperature, and density to the Z-factor as follows:

$$Z = 0.27\,\frac{P_{pr}}{T_{pr}\rho_r} \quad \text{(A8)}$$

where Ppr is pseudo-reduced pressure, Tpr pseudo-reduced temperature and $\rho_r$ is re duced density. In the above equation the reduced density is a function of eleven constants and is represented as follows:

$$\begin{aligned} F(\rho_r) = {} & A_1 + \frac{A_2}{T_{pr}} + \frac{A_3}{T_{pr}^3} + \frac{A_4}{T_{pr}^4} + \frac{A_5}{T_{pr}^5} + \rho_r\left(A_6 + \left(\frac{A_7}{T_{pr}}\right) + \left(\frac{A_8}{T_{pr}^2}\right)\right) \\ & + \rho_r^2\left(A_9\left(\frac{A_{10}}{T_{pr}^3}\right)\right) + \rho_r^3 A_{11}\left(\frac{1}{T_{pr}^4}\right) - \frac{P_{pr}}{T_{pr}} = 0 \end{aligned} \quad \text{(A9)}$$

where A1 to A11 are constants that were found by fitting the equation to 1500 data points from SK chart and are as follows:

A1 = 0.3265, A2 = −1.0700, A3 = −0.5339, A4 = 0.01569, A5 = −0.05165, A6 = 0.5475, A7 = −0.7361, A8 = 0.1844, A9 = 0.1056, A10 = 0.6134, A11 = 0.7210.

## References


1. EIA U.S. Energy Information Administration (EIA)–Widgets. Available online: https://www.eia.gov/opendata/embed.php?type=chart&series_id=NG.N9010US2.M&date_mode=all (accessed on 28 June 2020).
2. McCain, W.D.; Cawley, G. Reservoir-Fluid Property Correlations-State of the Art. *SPE Reserv. Eng.* **1991**, *6*, 266–272. [CrossRef]
3. Peng, D.Y.; Robinson, D.B. A New Two-Constant Equation of State. *Ind. Eng. Chem. Fundam.* **1976**, *15*, 59–64. [CrossRef]
4. Shateri, M.H.; Ghorbani, S.; Hemmati-Sarapardeh, A.; Mohammadi, A.H. Application of Wilcoxon Generalized Radial Basis Function Network for Prediction of Natural Gas Compressibility Factor. *J. Taiwan Inst. Chem. Eng.* **2015**, *50*, 131–141. [CrossRef]
5. Gaganis, V.; Homouz, D.; Maalouf, M.; Khoury, N.; Polychronopoulou, K. An Efficient Method to Predict Compressibility Factor of Natural Gas Streams. *Energies* **2019**, *12*, 2577. [CrossRef]
6. Kamari, A.; Mohammadi, A.H.; Ramjugernath, D. Petroleum Science and Technology Characterization of C 7+ Fraction Properties of Crude Oils and Gas-Condensates Using Data Driven Models Characterization of C 7þ Fraction Properties of Crude Oils and Gas-Condensates Using Data Driven Models. *Pet. Sci. Technol.* **2019**, *37*, 1516–1522. [CrossRef]
7. Ahmed, T.H. Comparative Study of Eight Equations of State for Predicting Hydrocarbon Volumetric Phase Behavior. *SPE Res. Eng.* **1988**, *3*, 337–348. [CrossRef]
8. Danesh, A. *PVT and Phase Behaviour of Petroleum Reservoir Fluids*, 1st ed.; Elsevier: Amsterdam, The Netherlands, 1998.
9. Azizi, N.; Behbahani, R.; Isazadeh, M.A. An Efficient Correlation for Calculating Compressibility Factor of Natural Gases. *J. Nat. Gas Chem.* **2010**, *19*, 642–645. [CrossRef]
10. Faraji, F.; Ugwu, J.O.; Chong, P.L. Modelling Two-Phase Z Factor of Gas Condensate Reservoirs: Application of Artificial Intelligence (AI). *J. Pet. Sci. Eng.* **2022**, *208*, 109787. [CrossRef]
11. Hemmati-Sarapardeh, A.; Hajirezaie, S.; Soltanian, M.R.; Mosavi, A.; Nabipour, N.; Shamshirband, S.; Chau, K.-W.; Hemmati–Sarapardeh, A.; Hajirezaie, S.; Soltanian, M.R.; et al. Modeling Natural Gas Compressibility Factor Using a Hybrid Group Method of Data Handling. *Eng. Appl. Comput. Fluid Mech.* **2019**, *14*, 27–37. [CrossRef]

12. Kamari, A.; Hemmati-Sarapardeh, A.; Mirabbasi, S.-M.; Nikookar, M.; Mohammadi, A.H. Prediction of Sour Gas Compressibility Factor Using an Intelligent Approach. *Fuel Process. Technol.* **2013**, *116*, 209–216. [CrossRef]
13. Whitson, C.; Brulé, M. *Phase Behavior*, 1st ed.; Society of Petroleum Engineers: Richardson, TX, USA, 2000; ISBN 1555630871.
14. Van der Waals, J. *Over de Continuiteit van Den Gas-En Vloeistoftoestand (On the Continuity of the Gas and Liquid State)*; University of Leiden: Leiden, The Netherland, 1873.
15. Soave, G. *Equilibrium Constants from a Modified Redkh-Kwong Equation of State*; Pergamon Press: New York, NY, USA, 1972; Volume 27.
16. Standing, M.B.; Katz, D.L. Density of Natural Gases. *Trans. AIME* **1942**, *146*, 140–149. [CrossRef]
17. Hall, K.R.; Yarborough, L. A New Equation of State for Z-Factor Calculations. *Oil Gas J.* **1973**, *71*, 82–92.
18. Dranchuk, P.M.; Abou-Kassem, J.H. Calculation of Z Factors for Natural Gases Using Equations of State. *J. Can. Pet. Technol.* **1974**, *14*, 34–36. [CrossRef]
19. Gopal, V. Gas Z-Factor Equation Developed for Computer. *Oil Gas J.* **1977**, *75*, 58–60.
20. Beggs, H.D.; Brill, J.R. Study of Two-Phase Flow in Inclined Pipes. *JPT J. Pet. Technol.* **1973**, *25*, 607–617. [CrossRef]
21. Kumar, N. *Compressibility Factors for Natural and Sour Reservoir Gases by Correlations and Cubic Equations of State*; Texas Tech University: Lubbock, TX, USA, 2004.
22. Heidaryan, E.; Moghadasi, J.; Salarabadi, A. A New and Reliable Model for Predicting Methane Viscosity at High Pressures and High Temperatures. *J. Nat. Gas Chem.* **2010**, *19*, 552–556. [CrossRef]
23. Hankinson, L.; Thomas, K.A.P. Predict Natural Gas Properties. *Hydrocarb. Process.* **1969**, *48*, 106–108.
24. Dranchuk, P.M.; Purvis, R.A.; Robinson, D.B. Computer Calculation Of Natural Gas Compressibility Factors Using The Standing And Katz Correlation. In Proceedings of the Annual Technical Meeting, Edmonton, AB, Canada, 7–11 May 1973; pp. 73–112.
25. Ogali, O.I.O.; Okoro, E.E.; Sanni, S.E.; John, I.T.; John, C.B. Advancing Z-Factor Prediction in Natural Gas Systems Using Machine Learning: A Review of Methods, Challenges, and Role in the Global Energy Transition. *Renew. Sustain. Energy Rev.* **2026**, *226*, 116254. [CrossRef]
26. Khosrojerdi, S.; Vakili, M.; Yahyaei, M.; Kalhor, K. Thermal Conductivity Modeling of Graphene Nanoplatelets/Deionized Water Nanofluid by MLP Neural Network and Theoretical Modeling Using Experimental Results. *Int. Commun. Heat Mass Transf.* **2016**, *74*, 11–17. [CrossRef]
27. Eslamimanesh, A.; Gharagheizi, F.; Illbeigi, M.; Mohammadi, A.H.; Fazlali, A.; Richon, D. Phase Equilibrium Modeling of Clathrate Hydrates of Methane, Carbon Dioxide, Nitrogen, and Hydrogen+water Soluble Organic Promoters Using Support Vector Machine Algorithm. *Fluid Phase Equilibria* **2012**, *316*, 34–45. [CrossRef]
28. Majidi, S.M.J.; Shokrollahi, A.; Arabloo, M.; Mahdikhani-Soleymanloo, R.; Masihi, M. Evolving an Accurate Model Based on Machine Learning Approach for Prediction of Dew-Point Pressure in Gas Condensate Reservoirs. *Chem. Eng. Res. Des.* **2014**, *92*, 891–902. [CrossRef]
29. Almsallti, M.; Alzubi, A.B.; Adegboye, O.R. Hybrid Metaheuristic Optimized Extreme Learning Machine for Sustainability Focused CO2 Emission Prediction Using Globalization-Driven Indicators. *Sustainability* **2025**, *17*, 6783. [CrossRef]
30. Soroush, E.; Mesbah, M.; Shokrollahi, A.; Rozyn, J.; Lee, M.; Kashiwao, T.; Bahadori, A. Evolving a Robust Modeling Tool for Prediction of Natural Gas Hydrate Formation Conditions. *J. Unconv. Oil Gas Resour.* **2015**, *12*, 45–55. [CrossRef]
31. Ahmadi, M.; Ebadi, M. Evolving Smart Approach for Determination Dew Point Pressure through Condensate Gas Reservoirs. *Fuel* **2014**, *117*, 1074–1084. [CrossRef]
32. Rostami, A.; Hemmati-Sarapardeh, A.; Shamshirband, S. Rigorous Prognostication of Natural Gas Viscosity: Smart Modeling and Comparative Study. *Fuel* **2018**, *222*, 766–778. [CrossRef]
33. Faraji, F.; Ugwu, J.O.; Chong, P.L. Development of a New Gas Condensate Viscosity Model Using Artificial Intelligence. *J. King Saud Univ.–Eng. Sci.* **2022**, *34*, 376–383. [CrossRef]
34. Al-Anazi, B.D.; Al Quraishi, A.A. New Correlation for Z-Factor Using Genetic Programming Technique. In Proceedings of the SPE Oil and Gas India Conference and Exhibition, Mumbai, India, 20–22 January 2010.
35. Kamyab, M.; Sampaio, J.H.B.; Qanbari, F.; Eustes, A.W. Using Artificial Neural Networks to Estimate the Z-Factor for Natural Hydrocarbon Gases. *J. Pet. Sci. Eng.* **2010**, *73*, 248–257. [CrossRef]
36. Sanjari, E.; Lay, E.N. An Accurate Empirical Correlation for Predicting Natural Gas Compressibility Factors. *J. Nat. Gas Chem.* **2012**, *21*, 184–188. [CrossRef]
37. Chamkalani, A.; Mae'soumi, A.; Sameni, A. An Intelligent Approach for Optimal Prediction of Gas Deviation Factor Using Particle Swarm Optimization and Genetic Algorithm. *J. Nat. Gas Sci. Eng.* **2013**, *14*, 132–143. [CrossRef]
38. Azizi, N.; Rezakazemi, M.; Zarei, M.M. An Intelligent Approach to Predict Gas Compressibility Factor Using Neural Network Model. *Neural Comput. Appl.* **2019**, *31*, 55–64. [CrossRef]
39. Shokir, E.M.E.M.; El-Awad, M.N.; Al-Quraishi, A.A.; Al-Mahdy, O.A. Compressibility Factor Model of Sweet, Sour, and Condensate Gases Using Genetic Programming. *Chem. Eng. Res. Des.* **2012**, *90*, 785–792. [CrossRef]

40. Hemmati-Sarapardeh, A.; Varamesh, A.; Husein, M.M.; Karan, K. On the Evaluation of the Viscosity of Nanofluid Systems: Modeling and Data Assessment. *Renew. Sustain. Energy Rev.* **2018**, *81*, 313–329. [CrossRef]
41. Yan, K.L.; Liu, H.; Sun, C.Y.; Ma, Q.L.; Chen, G.J.; Shen, D.J.; Xiao, X.J.; Wang, H.Y. Measurement and Calculation of Gas Compressibility Factor for Condensate Gas and Natural Gas under Pressure up to 116 MPa. *J. Chem. Thermodyn.* **2013**, *63*, 38–43. [CrossRef]
42. Buxton, T.S.; Campbell, J.M. Compressibility Factors for Lean Natural Gas-Carbon Dioxide Mixtures at High Pressure. *Soc. Pet. Eng. J.* **1967**, *7*, 80–86. [CrossRef]
43. Liu, H.; Sun, C.Y.; Yan, K.L.; Ma, Q.L.; Wang, J.; Chen, G.J.; Xiao, X.J.; Wang, H.Y.; Zheng, X.T.; Li, S. Phase Behavior and Compressibility Factor of Two China Gas Condensate Samples at Pressures up to 95MPa. *Fluid Phase Equilibria* **2013**, *337*, 363–369. [CrossRef]
44. Li, Q.; Guo, T.M. A Study on the Supercompressibility and Compressibility Factors of Natural Gas Mixtures. *J. Pet. Sci. Eng.* **1991**, *6*, 235–247. [CrossRef]
45. Sun, C.Y.; Liu, H.; Yan, K.L.; Ma, Q.L.; Liu, B.; Chen, G.J.; Xiao, X.J.; Wang, H.Y.; Zheng, X.T.; Li, S. Experiments and Modeling of Volumetric Properties and Phase Behavior for Condensate Gas under Ultra-High-Pressure Conditions. *Ind. Eng. Chem. Res.* **2012**, *51*, 6916–6925. [CrossRef]
46. Satter, A.; Campbell, J.M. Non-Ideal Behavior of Gases and Their Mixtures. *Soc. Pet. Eng. J.* **1963**, *3*, 333–347. [CrossRef]
47. Haykin, S. *Neural Networks—A Comprehensive Foundation*, 5th ed.; Pearson: Singapore, 2005.
48. Bishop, C. *Pattern Recognition and Machine Learning*, 1st ed.; Springer Nature: Singapore, 2006.
49. Faraji, F.; Santim, C.; Chong, P.L.; Hamad, F. Two-Phase Flow Pressure Drop Modelling in Horizontal Pipes with Different Diameters. *Nucl. Eng. Des.* **2022**, *395*, 111863. [CrossRef]
50. Faraji, F.; Ugwu, J.; Chong, P.L.; Nabhani, F. Modelling Viscosity of Liquid Dropout near Wellbore Region in Gas Condensate Reservoirs Using Modern Numerical Approaches. *J. Pet. Sci. Eng.* **2020**, *185*, 106604. [CrossRef]
51. Ivakhnenko, A.G. Polynomial Theory of Complex Systems. *IEEE Trans. Syst. Man Cybern.* **1971**, *1*, 364–378. [CrossRef]
52. Shariaty, S.; Khorsand Movaghar, M.R.; Vatandoost, A. A New Model for Estimating the Gas Compressibility Factor Using Group Method of Data Handling Algorithm (Case Study). *Asia-Pacific J. Chem. Eng.* **2019**, *14*, e2307. [CrossRef]
53. Koza, J.R. Genetic Programming as a Means for Programming Computers by Natural Selection. *Stat. Comput.* **1994**, *4*, 87–112. [CrossRef]
54. Najafi-Marghmaleki, A.; Tatar, A.; Barati-Harooni, A.; Arabloo, M.; Rafiee-Taghanaki, S.; Mohammadi, A.H. Reliable Modeling of Constant Volume Depletion (CVD) Behaviors in Gas Condensate Reservoirs. *Fuel* **2018**, *231*, 146–156. [CrossRef]
55. Saghafi, H.; Arabloo, M. Development of Genetic Programming (GP) Models for Gas Condensate Compressibility Factor Determination below Dew Point Pressure. *J. Pet. Sci. Eng.* **2018**, *171*, 890–904. [CrossRef]
56. Fabani, M.P.; Capossio, J.P.; Román, M.C.; Zhu, W.; Rodriguez, R.; Mazza, G. Producing Non-Traditional Flour from Watermelon Rind Pomace: Artificial Neural Network (ANN) Modeling of the Drying Process. *J. Environ. Manag.* **2021**, *281*, 111915. [CrossRef]
57. Ghanem, A.; Gouda, M.F.; Alharthy, R.D.; Desouky, S.M. Predicting the Compressibility Factor of Natural Gas by Using Statistical Modeling and Neural Network. *Energies* **2022**, *15*, 1807. [CrossRef]
58. Kanchev, N.; Stoyanov, N.; Milushev, G. Prediction of the Natural Gas Compressibility Factor by Using MLP and RBF Artificial Neural Networks. *Meas. Sci. Rev.* **2025**, *25*, 1–9. [CrossRef]
59. Elliott, J.R.; Lira, C.T. *Introductory Chemical Engineering Thermodynamics*, 2nd ed.; Pearson: London, UK, 2012.
60. Simon, R.; Briggs, J.E. Application of the Benedict–Webb–Rubin Equation of State to Hydrogen Sulfide–Hydrocarbon Mixtures. *AIChE J.* **1964**, *10*, 548–550. [CrossRef]