

Predicting Thermodynamic Properties of R1234yf/R152a Mixtures Using Machine Learning and PRSV Equation of state

Sadegh Sahraei

sahraei.s@lu.ac.ir

Department of Polymer Engineering, Faculty of Engineering, Lorestan University, Khorramabad, Iran

Fatemeh Karami

Lorestan University

Mohammad Mahdi Tafarroj

Lorestan University

Method Article

Keywords: Machine Learning, Refrigerant Mixtures, Compressibility Factor, Density, PRSV Equation of State

Posted Date: September 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10515541/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Predicting Thermodynamic Properties of R1234yf/R152a Mixtures Using Machine Learning and PRSV Equation of state

Sadegh Sahraei^{1,*}, Fatemeh Karami², Mohammad Mahdi Tafarroj²

¹ Department of Polymer Engineering, Faculty of Engineering, Lorestan University, Khorramabad, Iran

² Department of Mechanical Engineering, Faculty of Engineering, Lorestan University, Khorramabad, Iran

* Corresponding author, E-mail address: sahraei.s@lu.ac.ir

Abstract

The phase-down of high-GWP hydrofluorocarbons has driven interest in eco-friendly refrigerants such as R1234yf/R152a. Accurate prediction of their compressibility factor and density is essential for refrigeration system design. This study compares four machine learning models Gradient Boosting (GB), Random Forest (RF), Support Vector Regression with RBF kernel (SVR-RBF), and K-Nearest Neighbors (KNN) with the PRSV equation of state and the Tait equation for predicting these properties. Experimental PVTx data (145 points) were collected for R1234yf/R152a mixtures (mass fraction: 0.6–0.9, temperature: 288.15–328.15 K). GB achieved the highest accuracy for compressibility factor ($R^2 = 0.8493$, RMSE = 0.0207), while SVR-RBF excelled for density ($R^2 = 0.9996$, RMSE = 0.2727). The PRSV equation with quadratic temperature-dependent k_{ij} improved Z-factor predictions ($R^2 = 0.7429$) compared to constant k_{ij} ($R^2 = 0.7080$). For density, PRSV with exponential k_{ij} achieved $R^2 = 0.9971$, closely followed by the Tait equation ($R^2 = 0.9954$). Results demonstrate that ensemble ML models outperform classical equations of state for Z-factor prediction, while both data-driven and thermodynamic approaches yield near-perfect density predictions.

Keywords: Machine Learning, Refrigerant Mixtures, Compressibility Factor, Density, PRSV Equation of State.

1. Introduction

The escalating concern over global warming has triggered stringent regulatory actions regarding the use of hydrofluorocarbons (HFCs). Although HFCs possess zero ozone depletion potential (ODP), their high Global Warming Potential (GWP) significantly contributes to climate change [1, 2]. In response, the parties to the Montreal Protocol agreed in October 2016 to accelerate the phase-down of HFCs [3]. This decision has created an urgent demand for environmentally friendly refrigerants that meet both environmental and performance criteria. Traditional refrigerants like R134a, with a GWP of 1430, are no longer viable for many applications, necessitating the development and adoption of low-GWP alternatives [4, 5].

Among the emerging candidates, R1234yf (2,3,3,3-tetrafluoropropene) has gained considerable attention due to its zero ODP and extremely low GWP (<1.5) [6]. Numerous studies have evaluated R1234yf as a drop-in replacement for R134a in automotive air conditioning and heat pump systems [7-10]. Research indicates that while R1234yf offers comparable cooling capacity and energy efficiency, its performance characteristics differ slightly from R134a, requiring precise thermodynamic data for system optimization [11-13]. For instance, some studies report a slight reduction in cooling capacity and an increase in power consumption, leading to a marginal decrease in the coefficient of performance (COP). Despite these minor differences, R1234yf is widely regarded as the most suitable alternative for mobile air conditioning due to its low toxicity and flammability classification [12, 14-16]. To further enhance safety and performance, binary mixtures of R1234yf with other refrigerants, such as R152a (1,1-difluoroethane), have been investigated. R152a has a low GWP (140) and a short atmospheric lifetime but is classified as flammable [9, 17, 18]. Mixing it with the mildly flammable R1234yf can result in a non-flammable or less flammable mixture, improving safety without significantly compromising thermodynamic performance [19, 20]. Previous studies have measured the vapor-liquid equilibrium and PVTx properties of R1234yf/R152a mixtures, providing valuable experimental data [21]. However, modeling these properties accurately remains challenging due to the complex non-linear interactions between temperature, pressure, and composition.

Traditional approaches to modeling refrigerant properties rely on EoS, such as the virial equation or cubic equations, which require fitting complex coefficients to experimental data. While effective, these methods can be computationally intensive and may not always capture subtle non-linear trends in the data, especially in the near-critical region [22-24]. In recent years, machine learning (ML) has emerged as a powerful alternative for predicting thermodynamic properties. ML models can learn complex patterns from data without explicit physical assumptions, offering high accuracy and computational efficiency. Recent studies have demonstrated the superiority of ML over traditional EoS in predicting properties of refrigerant mixtures with minimal data requirements [25-27]. Recent studies have increasingly adopted machine learning techniques for refrigerant property prediction. Noushabadi et al. [28] applied multiple algorithms to model thermodynamic and transport properties across various refrigerant classes, achieving high predictive accuracy. He et al. [29] developed an attention-enhanced data-driven framework for predicting critical and boiling properties of fluorine/chlorine-based refrigerants, demonstrating the value of integrating molecular descriptors into ML architectures. Wang and Hu [30] combined graph neural networks with transfer learning to estimate normal boiling points and critical temperatures, improving generalization for novel refrigerants with scarce experimental data. More recently, Dikmen [31] compared linear regression, artificial neural networks, and gene expression programming for predicting thermodynamic properties of the low-GWP refrigerant R466A, concluding that non-linear methods consistently outperformed simple regression models. These investigations underscore the expanding role of data-driven approaches in refrigerant property estimation and support their extension to binary mixtures such as R1234yf/R152a.

This study employs both machine learning and classical thermodynamic approaches to predict the compressibility factor and density of R1234yf/R152a mixtures using experimental PVTx data. Four ML algorithms Gradient Boosting (GB), Random Forest (RF), Support Vector Regression with RBF kernel (SVR-RBF), and K-Nearest Neighbors (KNN) are trained and optimized through hyperparameter tuning with 5-fold cross-validation. Additionally, the PRSV (Peng-Robinson-Stryjek-Vera) equation of state with van der Waals mixing rules is applied, where binary interaction parameters (k_{ij}) are optimized using various temperature-dependent formulations including linear, quadratic, exponential, and power-law functions to enhance predictive accuracy. The Tait equation is also employed as an empirical correlation for density prediction. The performance of these data-driven and physically grounded models is systematically compared, offering a comprehensive evaluation of their capabilities for predicting the thermodynamic properties of this environmentally friendly refrigerant mixture.

2. Methodology

2.1. Data Collection and Preprocessing

The experimental dataset used in this study comprises 145 samples of PVTx properties for R1234yf/R152a mixtures, sourced from the experimental work of [21]. The data covers a mass fraction of R1234yf ranging from 0.600 to 0.900, with temperatures between 288.15 K and 328.15 K. The dataset includes three primary input features: temperature (T), pressure (P), and the mass fraction of R1234yf (w). The target variables for prediction are the compressibility factor (Z) and density (ρ). Data preprocessing involved checking for missing values, which were absent, and ensuring the consistency of the experimental measurements.

Pearson correlation heatmap (Figure 1) exhibits strong negative correlations of Z-factor with Pressure ($r = -0.864$) and Density ($r = -0.878$), confirming that higher pressure reduces compressibility as real gases deviate from ideal behavior. Density shows a very strong positive correlation with Pressure ($r = 0.987$), which is thermodynamically expected since density increases with pressure at constant temperature. This near-collinearity between the two input variables (Pressure and the target Density) is inherent to PVTx data and does not pose a multicollinearity issue for modeling, as Density is a target variable, not a predictor for Z-factor. Temperature displays a weak negative correlation with Z-factor ($r = -0.115$) and moderate positive correlations with Pressure ($r = 0.355$) and Density ($r = 0.295$), indicating its influence is partially indirect through pressure effects. Mass Fraction shows negligible correlation with all variables ($|r| < 0.13$), including both targets. However, this does not imply composition is unimportant—its effects on thermodynamic properties are inherently non-linear and are captured effectively by tree-based and kernel methods. Overall, Pressure emerges as the dominant linear driver of both target variables, while temperature and composition require non-linear modeling approaches to fully capture their contributions.

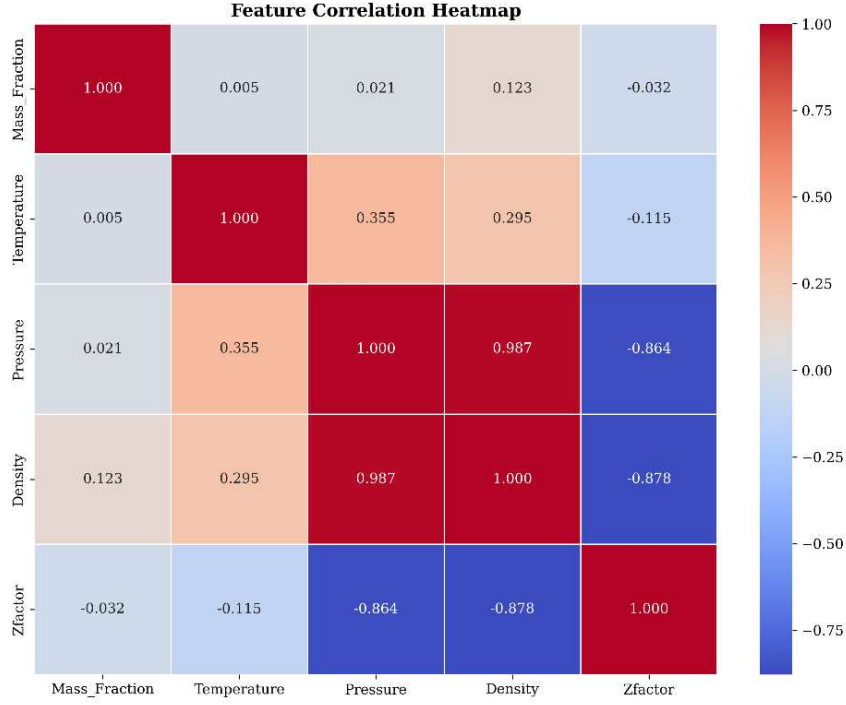


Fig. 1. Pearson correlation heatmap showing linear relationships among input features and Z-factor/Density.

2.2. Machine Learning Models

Four regression algorithms with distinct learning paradigms were selected to predict the compressibility factor and density of R1234yf/R152a mixtures. The choice of models was guided by their proven effectiveness for thermodynamic property prediction and their ability to capture both linear and non-linear relationships in PVTx data.

2.2.1 Support Vector Regression (SVR-RBF)

To capture non-linear relationships between thermodynamic variables (temperature, pressure, composition) and target properties (density and Z-factor), the SVR algorithm with a radial basis function kernel was employed. SVR identifies an optimal regression function by allowing data points within an ε -insensitive band, with penalties applied only for deviations exceeding this threshold. This mechanism enhances robustness against noise and prevents overfitting. The prediction function is expressed as:

$$f(x) = \sum_{i=1}^n (\alpha_i - \alpha_i^*) K(x_i, x) + b$$

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$$

where $K(x_i, x_j)$ is the RBF kernel, α_i and α_i^* are Lagrange multipliers, and b is the bias term. The three key hyperparameters the regularization parameter (C), the kernel coefficient (γ), and the

insensitive-loss parameter (ϵ) were optimized using grid search combined with k-fold cross-validation to minimize prediction error and ensure good generalization on unseen data.

2.2.2 Random Forest

Random Forest is an ensemble method that constructs multiple decision trees during training. Each tree is developed from a bootstrap sample, and at each node, a random subset of predictor variables is examined, reducing inter-tree correlation. This strategy of bootstrap aggregation and stochastic feature selection enhances generalization and mitigates overfitting. For regression tasks, the ensemble prediction is obtained by averaging individual tree outputs:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N T_i(x)$$

where N denotes the total count of trees within the forest, and $T_i(x)$ represents the output of the i -th decision tree corresponding to input vector x . In the current investigation, the principal hyperparameters governing the Random Forest architecture—specifically, the number of estimators, the maximum allowable depth of each tree, and the minimum sample size prerequisite for node bifurcation—were systematically optimized through a grid search procedure integrated with k-fold cross-validation. This optimization framework was implemented to attain superior predictive performance while simultaneously ensuring protection against overfitting.

2.2.3 Gradient Boosting

Gradient Boosting is a sequential ensemble method where each new tree corrects the residual errors of its predecessors. Unlike Random Forest, which builds trees independently, GB minimizes a loss function via gradient descent. The prediction after t iterations is:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$$

where $f_t(x_i)$ is the output of the newly added tree. The objective function combines a differentiable loss term with regularization to control model complexity:

$$L^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

where $\Omega(f_t)$ penalizes tree complexity to prevent overfitting. Key hyperparameters including number of estimators, learning rate, and maximum tree depth were optimized using grid search with 5-fold cross-validation. To evaluate the predictive performance of the proposed model, three statistical metrics were employed: the coefficient of determination R^2 , the root mean square error RMSE, and the mean absolute error MAE. These metrics are defined as follows:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

$$RMSE = \sqrt{\frac{1}{n} \sum (y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{n} \sum |y_i - \hat{y}_i|$$

where y_i , $\hat{y}_i$ and $\bar{y}_i$ represent the experimental, predicted values and mean of experimental data, respectively, and n is the number of data points.

2.2.4. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric instance-based learning algorithm that predicts target values by averaging the outputs of the k closest training samples in the feature space. The distance between samples is calculated using the Minkowski metric, where $p = 1$ corresponds to Manhattan distance and $p = 2$ to Euclidean distance. The prediction for a new sample is obtained by averaging the target values of its k nearest neighbors. Key hyperparameters—including the number of neighbors (k), distance metric (p), and weighting scheme (uniform or distance-based)—were optimized using grid search with 5-fold cross-validation to minimize prediction error and ensure robust generalization.

$$d(x_i, x_j) = \left(\sum_{m=1}^p |x_{i,m} - x_{j,m}|^p \right)^{1/p}$$

$$\hat{y}(x) = \frac{1}{k} \sum_{i=1}^k y_i$$

2.3. Model Training and Optimization

A structured training and optimization pipeline were implemented to guarantee model reliability and generalizability. A 5-fold cross-validation strategy was employed, where the 145 experimental data points were randomly partitioned into five subsets. During each round, four subsets served for training while the remaining one was used for validation, with the process repeated until every sample had been held out once. Final performance indicators were expressed as averages with corresponding standard deviations across all folds. Optimal hyperparameters were identified through grid search integrated with the same cross-validation procedure. The search ranges were customized for each algorithm: Gradient Boosting considered 100–300 estimators, tree depths of 3–5, and minimum split samples of 2–5; Random Forest explored 100–300 trees, depths from 5 to unlimited, and minimum split samples of 2–5; SVR-RBF examined regularization parameter C between 0.1 and 1000, epsilon values of 0.01–0.1, and gamma settings of scale, 0.01, and 0.1; KNN tested neighbor counts from 3 to 15, distance orders of 1 and 2, and weighting options of uniform and distance-based.

Table 1. Optimal hyperparameters of machine learning models for predicting Z-factor and Density.

Model	Target	Optimal Hyperparameters
-------	--------	-------------------------

Gradient Boosting	Density	Estimators = 300, Max Depth = 3, Learning Rate = 0.1, Min Samples Split = 5, Min Samples Leaf = 2
	Z-factor	Estimators = 300, Max Depth = 3, Learning Rate = 0.1, Min Samples Split = 5, Min Samples Leaf = 1
Random Forest	Density	Estimators = 200, Max Depth = None, Min Samples Split = 2, Min Samples Leaf = 1
	Z-factor	Estimators = 100, Max Depth = 10, Min Samples Split = 2, Min Samples Leaf = 1
SVR-RBF	Density	C = 200, Epsilon = 0.05, Gamma = 0.1
	Z-factor	C = 10, Epsilon = 0.01, Gamma = 0.01
KNN	Density	Neighbors = 3, p = 2, Weights = distance
	Z-factor	Neighbors = 5, p = 2, Weights = distance

Table 1 presents the optimal hyperparameters for each model. For Gradient Boosting, the Z-factor configuration used 300 estimators, a maximum depth of 3, a learning rate of 0.1, minimum samples per split of 5, and minimum samples per leaf of 1. For density, minimum samples per leaf increased to 2, indicating slightly greater regularization. For Random Forest, the optimal Z-factor configuration used 100 estimators with a maximum depth of 10, while density required 200 estimators with no depth limit. For SVR-RBF, the Z-factor configuration used C = 10, epsilon = 0.01, and gamma = 0.01, while density required higher values (C = 200, epsilon = 0.05, gamma = 0.1), suggesting a more complex decision boundary. For KNN, the optimal number of neighbors was 5 for Z-factor and 3 for density, both using Euclidean distance with distance-based weighting.

2.4. Peng-Robinson Equation of State

In addition to machine learning models, the Peng-Robinson-Stryjek-Vera (PRSV) equation of state was employed as a classical thermodynamic approach for predicting the compressibility factor and density of R1234yf/R152a mixtures. The PRSV EoS is an improved version of the original Peng-Robinson equation, specifically developed for polar compounds and refrigerant mixtures. The PRSV equation of state is expressed as:

$$P = \frac{RT}{V_m - b} - \frac{\alpha(T)}{V_m(V_m + b) + b(V_m - b)}$$

where P is the pressure, R is the universal gas constant, T is the temperature, V_m is the molar volume, $\alpha(T)$ is the energy parameter, and b is the co-volume parameter. For pure components, the parameters a_i and b_i are calculated using the critical properties and acentric factor:

$$\alpha_i(T) = 0.45724 \frac{R^2 T_{c,i}^2}{P_{c,i}} \left[1 + k_i \left(1 - \sqrt{\frac{T}{T_{c,i}}} \right)^2 \right] \quad b_i = 0.0778 \frac{RT_{c,i}}{P_{c,i}}$$

where $T_{c,i}$ and $P_{c,i}$ are the critical temperature and pressure of component i , and k_i is a function of the acentric factor ω_i :

$$k_i = k_{0,i} + k_{1,i}(1 + \sqrt{T_{r,i}})(0.7 - T_{r,i})$$

$$k_{0,i} = 0.378893 + 1.4897153\omega_i - 0.17131848\omega_i^2 + 0.0196554\omega_i^3$$

Table 2. Critical properties and acentric factors of R1234yf and R152a.

Property	R1234yf	R152a
T_c (K)	367.85	386.41
P_c (MPa)	3.382	4.517
ω	0.276	0.275

For mixtures, the van der Waals (vdW) mixing rules were applied:

$$a_m = \sum_i \sum_j x_i x_j a_{ij} \quad b_m = \sum_i x_i b_i$$

where x_i is the mole fraction of component i , and a_{ij} is defined using the binary interaction parameter k_{ij} :

$$a_{ij} = \sqrt{a_i a_j} (1 - k_{ij})$$

The binary interaction parameter k_{ij} was optimized using various temperature-dependent formulations, including constant, linear, quadratic, exponential, and power-law functions. The quadratic temperature-dependent model ($k_{ij}=k_0+k_1(T-288.15)+k_2(T-288.15)^2$) yielded the best performance for Z-factor prediction with $k_0=-0.1586$, $k_1=0.0090$, and $k_2=0.0001$, achieving an R^2 of 0.7429. For density, the exponential model ($k_{ij}=k_0+k_1\exp(-T/k_2)$) performed best with $k_0=0.3544$, $k_1=-0.5000$, and $k_2=289.6345$, achieving an R^2 of 0.9971. The performance of the PRSV EoS was evaluated using the same statistical metrics (R^2 , RMSE, MAE) as the machine learning models, enabling a direct comparison between the two approaches.

2.5 Tait Equation

In addition to the PRSV equation of state, the Tait equation [32] was employed as an empirical correlation for density prediction. The Tait equation is widely used for correlating liquid and dense fluid densities as a function of temperature and pressure due to its simple mathematical form and high accuracy for compressed fluids. The modified Tait equation used in this study is expressed as:

$$\rho(P, T, w) = a_0 + a_1 T + a_2 P + a_3 w + a_4 T \cdot P + a_5 P \cdot w$$

where ρ is the density (kg/m^3), P is the pressure (kPa), T is the temperature (K), w is the mass fraction of R1234yf, and a_0 through a_5 are adjustable parameters determined by nonlinear regression. The six parameters were optimized using the training dataset by minimizing the sum of squared errors between experimental and predicted density values. The optimization was performed using the Levenberg-Marquardt algorithm. The optimal parameters obtained were:

$$\begin{aligned} a_0 &= 19.6374, & a_1 &= -0.0632, & a_2 &= 0.0177 \\ a_3 &= -3.0129, & a_4 &= 1.28 \times 10^{-5}, & a_5 &= 0.0346 \end{aligned}$$

The Tait equation offers several advantages: it requires no iterative solution, has a simple polynomial form, and provides fast predictions. However, as an empirical correlation, it lacks the theoretical foundation of cubic equations of state and may not extrapolate reliably beyond the calibrated range.

3. Results and Discussion

3.1. Performance Comparison for Compressibility Factor

Table 3 and Figure 2 compare the predictive performance of machine learning models and the PRSV equation of state for the compressibility factor. Gradient Boosting achieved the highest accuracy ($R^2 = 0.8493$, $\text{RMSE} = 0.0207$), with a cross-validation score of 0.8303 ± 0.0770 confirming robust generalization. Its sequential error-correction mechanism, which iteratively focuses on difficult-to-predict samples, proved particularly effective for capturing the sharp non-linear variations in Z -factor across the operating pressure range. SVR-RBF performed comparably ($R^2 = 0.8389$), though exhibited increased scatter at $Z < 0.85$. This degradation at lower Z -values can be attributed to the regularization parameter ($C = 10$), which prioritizes global generalization over local fit, causing the model to smooth out rapid changes in the compressibility factor near the saturation boundary. Random Forest showed marginally lower accuracy ($R^2 = 0.8348$), with a notable deviation near $Z \approx 0.77$. This outlier stems from the inherent limitation of tree-based methods: they partition the feature space into discrete regions and average predictions within each partition, making them less effective for extrapolating beyond the training domain where few data points exist. KNN achieved an R^2 of 0.7460, performing comparably to the best PRSV variant. As a distance-based method, KNN relies on local similarity rather than learning underlying functional relationships, which limits its ability to capture the global non-linear trends in Z -factor across the entire pressure-temperature-composition space.

Figure 4 depicts the sensitivity of Gradient Boosting performance to variations in the learning rate. As shown, test R^2 increased sharply for learning rates between 0.001 and 0.01, reflecting an underfitting regime in which the step size was too small for the ensemble to adequately capture the underlying data patterns within the fixed number of boosting iterations. Beyond this range, R^2 stabilized within a plateau of approximately 0.83–0.86, with the learning rate of 0.1 selected as the optimal operating point, offering a favorable balance between predictive accuracy and training stability. At higher learning rates (above 0.3), performance became increasingly unstable,

exhibiting pronounced oscillations and a downward trend, consistent with the model overshooting optimal gradient steps and losing generalization capacity. These findings indicate that moderate learning rates provide the most reliable trade-off between convergence speed and model robustness. Among thermodynamic approaches, the PRSV equation with quadratic temperature-dependent k_{ij} achieved the best result ($R^2 = 0.7429$), while the constant k_{ij} formulation performed poorest ($R^2 = 0.7080$). The 5% improvement from temperature-dependent k_{ij} demonstrates that incorporating physical temperature effects into the binary interaction parameter enhances predictive capability. However, the persistent performance gap between PRSV and machine learning models reflects the inherent limitations of cubic equations of state: they rely on simplified van der Waals mixing rules and a single adjustable parameter to represent complex molecular interactions in polar refrigerant mixtures, which is insufficient for capturing the full non-ideal behavior. A systematic underprediction of the data point near $Z \approx 1.03$ was observed across all models, indicating insufficient training representation in the near-ideal gas region where the compressibility factor approaches unity. This observation underscores the importance of comprehensive data coverage for both data-driven and physics-based modeling approaches.

Figure 3 presents the residual analysis for the Gradient Boosting model. The residuals are centered near zero (mean = 0.0047) with a standard deviation of 0.0202 and MAE of 0.0104, confirming generally accurate predictions. Most residuals fall within ± 0.02 with no systematic trend, indicating negligible bias and homogeneous variance. However, three outliers are evident—particularly a residual of +0.087 near a predicted Z of 0.95—suggesting that the model underperforms for a small subset of boundary or sparsely sampled conditions. The histogram reveals a sharp, right-skewed peak concentrated between -0.01 and $+0.01$, but a thin tail extending toward $+0.08$ captures the few large positive errors. These outlier cases, while rare, contribute disproportionately to the overall error and could be addressed through targeted data augmentation or additional feature engineering in the corresponding operating region.

Table 3. The performance of the four models for the compressibility factor

Model	R^2	RMSE	MAE
Gradient Boosting	0.8493	0.0207	0.0104
SVR	0.8389	0.0214	0.0151
Random Forest	0.8348	0.0217	0.0135
KNN	0.7460	0.0269	0.0185
PRSV (Quadratic $k_{ij}(T)$)	0.7429	0.0271	0.0153
PRSV (Linear $k_{ij}(T)$)	0.7395	0.0273	0.0155
PRSV ($k_{ij}(T,P)$)	0.7393	0.0273	0.0154
PRSV (Power law $k_{ij}(T)$)	0.7186	0.0283	0.0166
PRSV (Exponential $k_{ij}(T)$)	0.7114	0.0287	0.0172
PRSV (Constant k_{ij})	0.7080	0.0289	0.0174

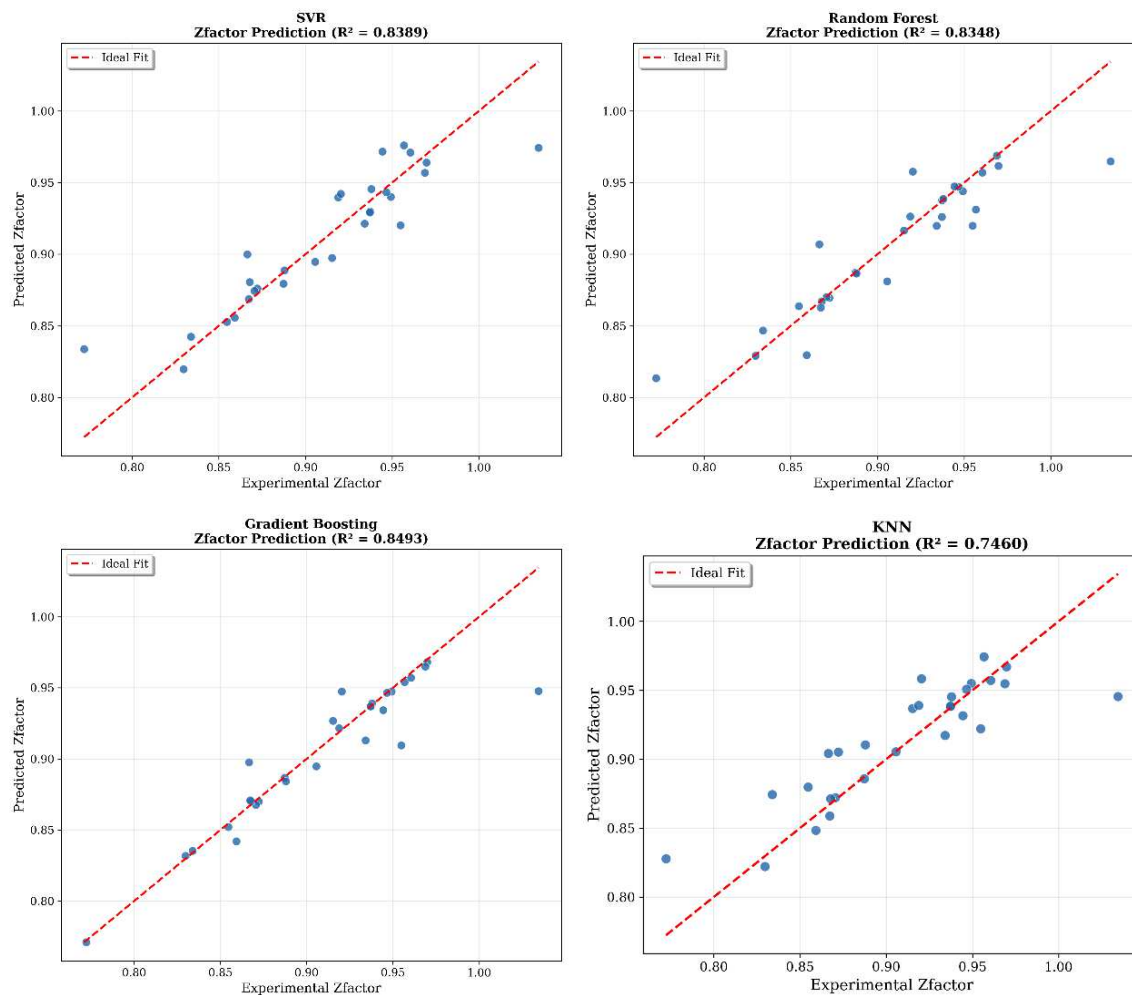


Fig. 2. Actual vs. Predicted Z-factor for all models

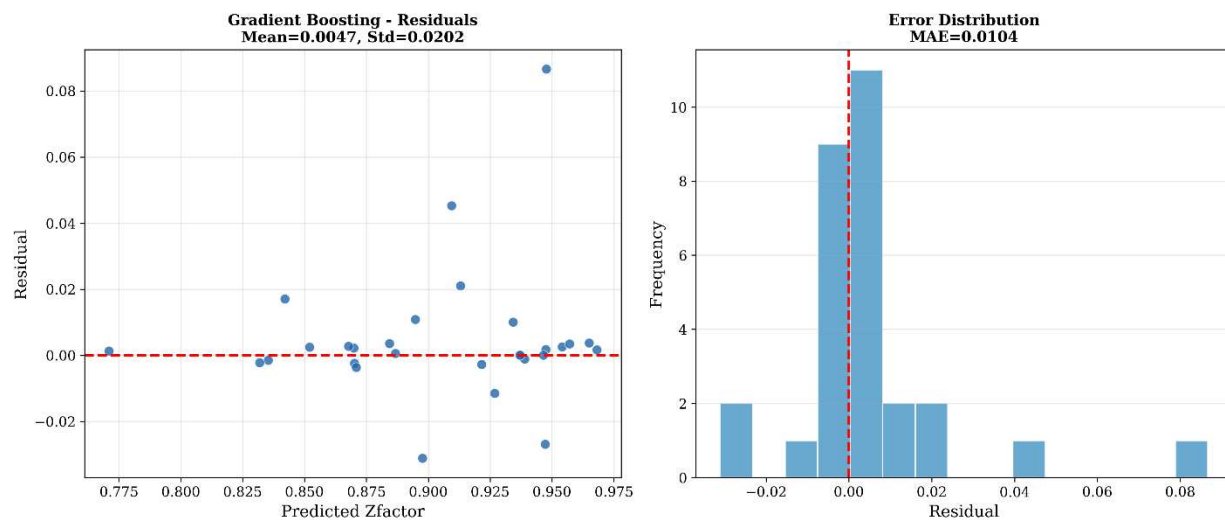


Fig. 3. Residual Plot of Gradient Boosting for Compressibility Factor Prediction

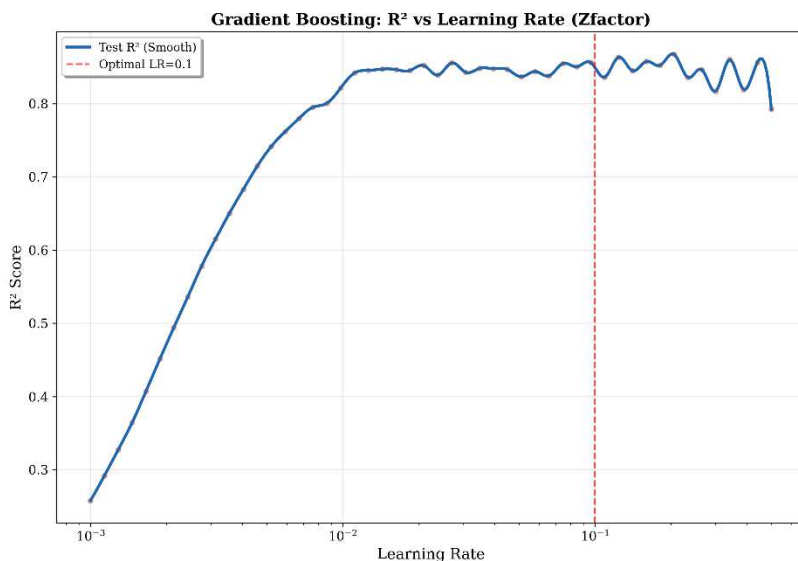


Fig. 4. Gradient Boosting: R^2 vs. Learning Rate, showing optimal performance at learning rate = 0.1.

3.2. Performance Comparison for Density

Table 4 summarizes the density prediction results, which contrast sharply with those for the compressibility factor. SVR-RBF achieved near-perfect accuracy ($R^2 = 0.9996$, RMSE = 0.2727, MAE = 0.2074), with a cross-validation score of 0.9984 ± 0.0019 confirming exceptional stability. The optimal parameters ($C = 200$, epsilon = 0.05, gamma = 0.1) enabled the RBF kernel to capture the underlying physical relationships without overfitting. Figure 6 illustrates how changes in the regularization parameter C affect the performance of the SVR model with an RBF kernel. For very low C values (0.1–1), the test R^2 remained low or negative, indicating underfitting caused by excessive regularization that restricted the model's ability to capture density patterns. As C increased beyond 1, the R^2 rose rapidly and stabilized near 1.0 around $C = 10$, with the peak performance observed at $C = 200$. Beyond this point, raising C further to 1000 yielded no noticeable improvement in R^2 , while adding unnecessary computational expense. These observations confirm that $C = 200$ provides the optimal trade-off between model flexibility and generalization for this dataset. Among thermodynamic models, the PRSV equation with exponential temperature-dependent k_{ij} performed best ($R^2 = 0.9971$, RMSE = 0.7057), closely followed by the constant k_{ij} formulation ($R^2 = 0.9970$). The Tait equation achieved an R^2 of 0.9954 with an RMSE of 0.8934. While slightly less accurate than the best PRSV variants, the Tait equation offers a simpler functional form with only six adjustable parameters, making it computationally efficient and easy to implement. KNN achieved an R^2 of 0.9549, performing comparably to Random Forest but significantly lower than SVR-RBF and the thermodynamic

models. As a local method, KNN struggles with the smooth, continuous nature of density variations, where global patterns dominate over local neighborhoods.

Notably, the top thermodynamic models performed nearly as well as SVR-RBF, with R^2 values exceeding 0.997. This contrasts sharply with the Z-factor results, where machine learning models significantly outperformed the PRSV equation. The strong performance of classical correlations for density suggests that density exhibits a more linear and physically predictable relationship with temperature, pressure, and composition compared to the compressibility factor. Gradient Boosting ($R^2 = 0.9711$) and Random Forest ($R^2 = 0.9564$) underperformed relative to both SVR-RBF and the best thermodynamic models. This likely reflects the predominantly linear nature of density behavior, where the complex non-linear structures of ensemble methods introduce unnecessary variance rather than capturing meaningful patterns. The Tait equation's competitive performance further supports this interpretation, as its simple polynomial form effectively captures the smooth, continuous density variations across the operating range.

Table 4. The performance of the four models for Density

Model	R^2	RMSE	MAE
SVR-RBF	0.9996	0.2727	0.2074
PRSV (Exponential k_{ij})	0.9971	0.7057	0.4476
PRSV (Constant k_{ij})	0.9970	0.7155	0.4581
PRSV (Power law k_{ij})	0.9969	0.7280	0.4509
PRSV (Quadratic k_{ij})	0.9962	0.8089	0.4357
PRSV ($k_{ij}(T,P)$)	0.9962	0.8089	0.4443
PRSV (Linear k_{ij})	0.9961	0.8196	0.4480
Tait Equation	0.9954	0.8934	0.6511
Gradient Boosting	0.9711	2.2340	1.4180
Random Forest	0.9564	2.7456	1.8000
KNN	0.9549	2.7906	2.1734

Figure 5 presents the overfitting diagnosis for the SVR-RBF model. The learning curve shows rapid convergence of both training and cross-validation R^2 scores to near-unity values beyond approximately 20 training samples, with the initial confidence interval narrowing substantially as data increases. This indicates stable generalization with a sufficiently sized training set. The performance comparison confirms near-identical R^2 values between training (0.9999) and test (0.9996) sets, yielding a negligible gap of 0.0004 well below typical overfitting thresholds. However, error-based metrics reveal a more meaningful divergence: RMSE increases from 0.1188 (training) to 0.2727 (test), and MAE from 0.0787 to 0.2074, representing roughly a twofold increase. This pattern suggests that while the model captures the overall variance structure exceptionally well, a small number of test observations contribute disproportionately to the error metrics. The residual analysis supports this interpretation: training residuals (Std = 0.1185) cluster tightly around zero, while test residuals (Std = 0.2696) exhibit greater dispersion with notable

outliers near predicted densities of 38 (+0.82) and 58 (−0.62). These localized deviations indicate that test-set errors are not uniformly distributed but concentrated in a few poorly predicted observations, likely corresponding to underrepresented regions of the density distribution. Despite these localized errors, the minimal R^2 gap confirms no significant overfitting.



Fig. 5. Overfitting diagnosis for the SVR-RBF model applied to density prediction: learning curve, training vs. test performance comparison, residual analysis, and performance gap summary.

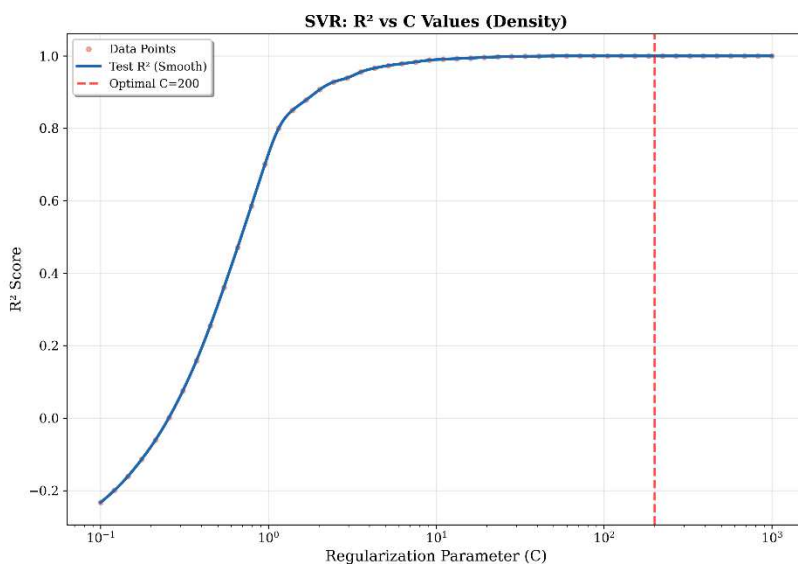


Fig. 6. R^2 vs. Regularization Parameter C , showing optimal performance at $C = 200$ for SVR model.

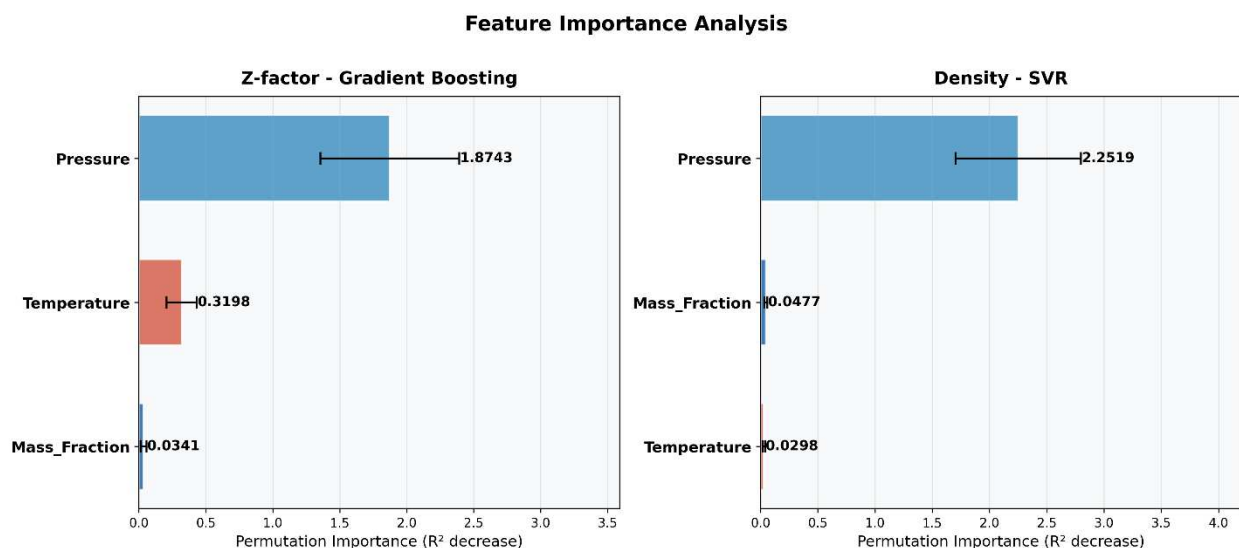


Fig. 7. Permutation feature importance analysis for the top-performing models

To improve the physical interpretability of the machine learning models and explain their advantage over classical equations of state, permutation feature importance was computed for the best-performing models (Gradient Boosting for Z-factor and SVR-RBF for density). Figure 7 shows that pressure dominates both predictions, with importance scores of 1.8743 for Z-factor and 2.2519 for density, far exceeding those of temperature and composition. This finding aligns with the Pearson correlation analysis (Figure 1), where pressure exhibited the strongest linear relationships with both targets ($r = -0.864$ with Z-factor and $r = 0.987$ with density).

From a thermodynamic perspective, this dominance is physically expected. Within the studied range, increasing pressure intensifies real-gas deviations, reducing the compressibility factor while sharply increasing density due to decreasing molar volume. Temperature plays a secondary role, with importance scores of 0.3198 for Z-factor but only 0.0298 for density, reflecting its greater influence on compressibility near phase boundaries through molecular kinetic energy effects. The mass fraction of R1234yf contributes minimally in linear terms (0.0341 and 0.0477, respectively), yet its non-linear interactions with pressure and temperature are effectively captured by the ensemble and kernel methods. The physical consistency of the ML models where pressure emerges as the primary driver validates their reliability for refrigeration system design and highlights opportunities for targeted sensor placement and process optimization.

4. Conclusion

This study compared machine learning models (Gradient Boosting, Random Forest, SVR-RBF) with classical thermodynamic correlations (PRSV with various kij formulations, and the Tait equation) for predicting the compressibility factor and density of R1234yf/R152a mixtures, based on 145 experimental data points. Gradient Boosting achieved the best Z-factor prediction ($R^2 = 0.8493$), clearly outperforming the top PRSV variant with quadratic kij ($R^2 = 0.7429$), underscoring the limits of cubic equations of state in modeling Z-factor's nonlinear behavior. For density, results were uniformly strong: SVR-RBF reached near-perfect accuracy ($R^2 = 0.9996$), closely followed by PRSV with exponential kij ($R^2 = 0.9971$) and the Tait equation ($R^2 = 0.9954$), reflecting density's more linear and physically predictable nature. Temperature-dependent kij improved PRSV's Z-factor accuracy by about 5% over the constant-kij model, though still falling short of ensemble ML performance. Overall, machine learning offers superior accuracy for complex nonlinear properties like Z-factor, while both data-driven and physics-based methods remain competitive for density. Future work should explore deep learning or physics-informed neural networks to enhance predictive robustness across broader conditions.

Nomenclature

CV	Cross-Validation
EoS	Equation of State
GB	Gradient Boosting
GWP	Global Warming Potential
HFC	Hydrofluorocarbon
KNN	K-Nearest Neighbors
LR	Linear Regression
MAE	Mean Absolute Error

ML Machine Learning
ODP Ozone Depletion Potential
PR Peng-Robinson
PRSV Peng-Robinson-Stryjek-Vera
R² Coefficient of Determination
RBF Radial Basis Function
RF Random Forest
RMSE Root Mean Square Error
SVR Support Vector Regression
vdW van der Waals

Declaration of Competing Interest

Funding: The authors declare that no funds, grants, or other financial support were received during the preparation of this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Data Availability Statement

The experimental data used in this study were extracted from the published work of [21]. The compiled dataset comprising 145 PVTx data points for R1234yf/R152a mixtures is available from the corresponding author upon reasonable request.

References

1. Salierno, G., *On the chemical pathways influencing the effective global warming potential of commercial hydrofluoroolefin gases*. ChemSusChem, 2024. **17**(17): p. e202400280.
2. Di Filippo, R., O.S. Bursi, and R. Di Maggio, *Global warming and ozone depletion potentials caused by emissions from HFC and CFC banks due to structural damage*. Energy and buildings, 2022. **273**: p. 112385.
3. Velders, G.J., et al., *Projections of hydrofluorocarbon (HFC) emissions and the resulting global warming based on recent trends in observed abundances and current policies*. Atmospheric Chemistry and Physics, 2022. **22**(9): p. 6087-6101.
4. Tsvetkov, O., et al. *Alternative refrigerants with low global warming potential for refrigeration and air-conditioning industries*. in *IOP Conference Series: Materials Science and Engineering*. 2020. IOP Publishing.
5. Heredia-Aricapa, Y., et al., *Overview of low GWP mixtures for the replacement of HFC refrigerants: R134a, R404A and R410A*. International Journal of Refrigeration, 2020. **111**: p. 113-123.
6. Sinche Chele, F., et al., *Critical Literature Review of Low Global Warming Potential (GWP) Refrigerants and their Environmental Impact*. 2024.

7. Aral, M.C., M. Suhermanto, and M. Hosoz, *Performance evaluation of an automotive air conditioning and heat pump system using R1234yf and R134a*. Science and Technology for the Built Environment, 2021. **27**(1): p. 44-60.
8. Hmood, K.S., et al., *Drop-in and retrofit refrigerants as replacement possibilities of R134a in domestic/commercial refrigeration and automobile air conditioner applications*. Journal of Thermal Engineering, 2021. **7**(7): p. 1815-1835.
9. Li, H. and K. Tang, *A comprehensive study of drop-in alternative mixtures for R134a in a mobile air-conditioning system*. Applied Thermal Engineering, 2022. **203**: p. 117914.
10. Meng, Z., et al., *Performance of R1234yf and R513A as alternatives to R134A in automotive air conditioning systems in winter*. Thermal Science, 2023. **27**(3 Part A): p. 1937-1946.
11. Hussain, Z., N. You, and H. Cho, *Performance and energy evaluation of R450A, R1234yf, R513A, R516A, R152A, and R1234ze (E) as alternatives to R134A in an air conditioner*. Applied Thermal Engineering, 2025. **271**: p. 126319.
12. Sharif, M., et al. *R1234yf vs R134a in automotive air conditioning system: A comparison of the performance*. in *IOP Conference Series: Materials Science and Engineering*. 2020. IOP Publishing.
13. Kuwar, Y.V. and G. Narasimham, *Performance of automotive air conditioning system with R134a and R1234yf*. International Journal of Air-Conditioning and Refrigeration, 2020. **28**(02): p. 2050013.
14. de Paula, C.H., et al., *Thermo-economic and environmental analysis of a small capacity vapor compression refrigeration system using R290, R1234yf, and R600a*. International Journal of Refrigeration, 2020. **118**: p. 250-260.
15. Alkan, A., A. Kolip, and M. Hosoz, *Energetic and exergetic performance comparison of an experimental automotive air conditioning system using refrigerants R1234yf and R134a*. Journal of Thermal Engineering, 2021. **7**(5): p. 1163-1173.
16. Patel, B. and A. Parekh, *Energy, exergy and entropy analysis with R1234yf as an alternate refrigerant to R134a of automobile air conditioning system*. Journal of Thermal Engineering, 2024. **10**(1): p. 101-114.
17. Kang, Y., et al., *Feasibility analysis on R1234yf/R152a replacing R134a as direct cooling refrigerant for electric vehicles*. Journal of Mechanical Science and Technology, 2024. **38**(5): p. 2661-2672.
18. Maeng, H., et al., *Energy and environmental performance of vapor injection heat pumps using R134a, R152a, and R1234yf under various injection conditions*. Energy, 2023. **280**: p. 128265.
19. Feng, B., Z. Yang, and H. Liu, *Research on inhibition of flame retardants on flammability of R1234yf*. International Journal of Refrigeration, 2020. **118**: p. 302-310.
20. Yang, Z., et al., *Analysis of lower GWP and flammable alternative refrigerants*. International Journal of Refrigeration, 2021. **126**: p. 12-22.
21. Xu, X., et al., *Experimental study on PVTx properties of a binary R1234yf/R152a refrigerant*. Journal of Chemical & Engineering Data, 2022. **67**(10): p. 2913-2919.
22. Kolobaev, V., et al., *Methodology for constructing the equation of state and thermodynamic tables for a new generation refrigerant*. Measurement Techniques, 2021. **64**(2): p. 86-93.
23. Su, W., et al., *Vapor-liquid equilibrium prediction of refrigerant mixtures with Peng–Robinson equation of state and binary interaction parameters calculated through group contribution model*. International Journal of Thermophysics, 2020. **41**(2): p. 14.
24. Rane, S., et al., *Analysis of real gas equation of state for CFD modelling of twin screw expanders with R245fa, R290, R1336mzz (Z) and R1233zd (E)*. International Journal of Refrigeration, 2021. **121**: p. 313-326.
25. Wan, H., et al., *Machine-learning-based compressor models: A case study for variable refrigerant flow systems*. International Journal of Refrigeration, 2021. **123**: p. 23-33.

26. Pence, I., et al., *Evaluation of machine learning approaches for estimating thermodynamic properties of new generation refrigerant R513A*. Sustainable Energy Technologies and Assessments, 2023. **55**: p. 102973.
27. Argirusis, N., et al., *A Machine Learning Approach for the Classification of Refrigerant Gases*. Applied Sciences, 2024. **14**(14): p. 6230.
28. Noushabadi, A.S., et al., *Insights into modelling and evaluation of thermodynamic and transport properties of refrigerants using machine-learning methods*. Energy, 2023. **262**: p. 125099.
29. He, Y., et al., *Data-driven approach augmented by attention mechanism in critical and boiling thermophysical properties prediction of fluorine/chlorine-based refrigerants*. Energy, 2024. **306**: p. 132490.
30. Wang, G. and P. Hu, *Prediction of normal boiling point and critical temperature of refrigerants by graph neural network and transfer learning*. International Journal of Refrigeration, 2023. **151**: p. 97-104.
31. Dikmen, E., *A comparative study of thermodynamic properties of R466A using linear regression, artificial neural network and gene expression programming*. Journal of Thermal Analysis and Calorimetry, 2024. **149**(21): p. 12265-12283.
32. Dymond, J. and R. Malhotra, *The Tait equation: 100 years on*. International journal of thermophysics, 1988. **9**(6): p. 941-951.