

Article

Predicting CO₂ Solubility in Brine for Carbon Storage with a Hybrid Machine Learning Framework Optimized by Ant Colony Algorithm

Seyed Hossein Hashemi , Farshid Torabi * and Sepideh Palizdan

Faculty of Engineering and Applied Science, University of Regina, Regina, SK S4S 0A2, Canada; shn081@uregina.ca (S.H.H.); spr019@uregina.ca (S.P.)

* Correspondence: farshid.torabi@uregina.ca

Abstract

Predicting carbon dioxide (CO₂) solubility in brine is critical for carbon capture and storage. This study employs the Ant Colony Optimization (ACO) algorithm to enhance the predictive accuracy of four machine learning models: Neural Network (NN), Decision Tree (DT), Support Vector Regression (SVR), and Gradient Boosting Machine (GBM). The models were trained and validated on a mineral compound dataset. Performance was evaluated using the coefficient of determination (R^2) and error metrics including RMSE and MAE. The GBM model achieved the highest test accuracy ($R^2 = 0.986$) with low errors (RMSE = 0.0478, MAE = 0.0362), demonstrating superior ability to model complex, non-linear relationships with minimal overfitting. The optimized NN, featuring three layers and fifteen neurons, delivered strong performance ($R^2 = 0.930$) with balanced errors across datasets. The DT model offered excellent interpretability and a strong test score ($R^2 = 0.912$), while the SVR model provided robust generalization ($R^2 = 0.889$). The results indicate that ACO is an effective tool for hyperparameter tuning across diverse model architectures. For maximum accuracy, GBM is recommended, whereas DT is ideal when interpretability is required. The NN presents a strong middle-ground option with competitive accuracy. This comparative framework assists in selecting the optimal model based on specific project priorities of accuracy, transparency, or computational efficiency for geochemical forecasting.

Keywords: CO₂ solubility; CO₂ storage; Ant Colony Algorithm; machine learning algorithm; GBM model



Academic Editors: Silvio José Gumiere and Hossein Bonakdari

Received: 13 February 2026

Revised: 6 March 2026

Accepted: 10 March 2026

Published: 11 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Rising atmospheric CO₂ from industrial activity drives climate change, necessitating effective carbon capture and storage solutions. Saline aquifers are a promising method for high-capacity, long-term CO₂ sequestration [1,2]. This process is driven by the dissolution of CO₂ into brine and the subsequent geochemical reactions that lead to permanent mineral trapping [3]. The dissolution of CO₂ in saline aquifers increases brine density by 0.1–1%, initiating natural convection that further enhances CO₂ dissolution through gravitational mixing [4]. The solubility of CO₂ is primarily controlled by reservoir temperature, pressure, and salinity [5].

Determining CO₂ solubility is essential for predicting its behavior and migration within geological storage sites, such as saline aquifers. While experimental measurements are often time-consuming and costly, machine learning has emerged as an ideal and powerful tool to accurately and efficiently predict CO₂ solubility in brine. In recent years,

researchers have increasingly adopted machine learning algorithms to address this challenge. Hashemi et al. [6] employed a Gaussian Process Regression (GPR) model optimized by the Grey Wolf Optimizer (GWO) to achieve highly accurate predictions of CO₂ solubility in brine. Their physically constrained machine learning framework demonstrates significant potential for optimizing CO₂ injection strategies in both carbon storage and enhanced oil recovery applications. Davoodi et al. [7] developed hybrid Long Short-Term Memory (LSTM) models optimized with metaheuristic algorithms, where the LSTM-COA model achieved the highest accuracy and robustness in predicting CO₂ solubility in diverse brine systems for carbon storage. Wei et al. [8] developed an accurate fusion model combining BPNN, GRNN, and XGBoost algorithms to predict CO₂ and H₂S solubility in brine, achieving superior performance over previous methods for applications in carbon storage and sour gas management. Yang et al. [9] developed an accurate Artificial Neural Network (ANN) model that effectively predicts CO₂ solubility in both pure water and brine, highlighting distinct dissolution behaviors between the two fluid systems for CCUS applications. Bhattacharjee et al. [10] demonstrated that an Extreme Gradient Boosting machine learning model provides a rapid and accurate method for estimating CO₂ fugacity coefficients, which subsequently enabled precise calculations of CO₂ solubility in saline solutions. Sadeghi et al. [11] successfully developed both thermodynamic and neural network models for predicting CO₂ solubility in NaCl brine, demonstrating that the optimized neural network offered comparable accuracy to the established thermodynamic approach for geological sequestration applications. Zou et al. [12] developed a Cascade Forward Neural Network optimized with the Levenberg–Marquardt algorithm (CFNN-LM) to accurately predict CO₂ solubility in multi-component brines, identifying pressure as the most influential parameter and ranking the salting-out effects of various salts for CCS applications.

In this study, to address the critical need for predicting carbon dioxide solubility in brine for carbon storage, we propose a novel hybrid approach combining machine learning with metaheuristic optimization. A review of the existing literature (Refs. [6–12]) reveals that previous studies have employed various machine learning and optimization approaches, including GPR-GWO [6], LSTM-COA [7], BPNN-GRNN-XGBoost fusion [8], ANN [9], XGBoost [10], thermodynamic and neural network hybrids [11], and CFNN-LM [12], to predict CO₂ solubility in brine systems. Despite their valuable contributions, these studies are generally limited in two key aspects: (1) they focus on a narrow range of machine learning models or specific optimization algorithms, and (2) the ionic composition of the brine systems considered is often restricted to common ions (e.g., Na⁺, K⁺, Ca²⁺, Mg²⁺), with limited attention to more complex and diverse ionic species. In contrast, the current study introduces a novel and comprehensive framework by evaluating four distinct machine learning algorithms, namely Neural Networks (NN), Decision Trees (DT), Support Vector Regression (SVR), and Gradient Boosting Machines (GBM), each optimized using the Ant Colony Optimization (ACO) algorithm, which has rarely been applied in this domain. Furthermore, the dataset employed in this work encompasses a wide variety of ions, thereby significantly expanding the chemical complexity and applicability of the models to real-world CCUS scenarios. This novel combination of diverse learning paradigms, a unique optimization strategy, and an expanded ionic database positions the current study as a meaningful advancement in the predictive modeling of CO₂ solubility for carbon capture, utilization, and storage applications.

2. Data and Methods

2.1. Hybrid Machine Learning–ACO Framework for CO₂ Solubility Prediction

The primary objective of this study is to develop a hybrid predictive framework that integrates four machine learning models—Neural Network (NN), Decision Tree (DT), Gradient Boosting Machine (GBM), and Support Vector Regression (SVR)—with the Ant Colony Optimization (ACO) metaheuristic for systematic hyperparameter tuning. The proposed framework is designed to accurately predict CO₂ solubility in brine systems while maintaining model robustness and generalization capability.

2.1.1. Unified ACO-Based Hyperparameter Optimization Procedure

To ensure methodological consistency and avoid redundancy, a unified ACO optimization workflow was implemented for all machine learning models. The overall procedure consists of the following stages:

1. Data Preparation and Partitioning

The dataset, consisting of the feature matrix (X) and target solubility vector (Y), was randomly shuffled using a fixed random seed to guarantee reproducibility. The data were divided into three subsets: 70% for training, 15% for validation, and 15% for testing. The training set was used for model fitting, the validation set for hyperparameter evaluation during optimization, and the test set for final unbiased performance assessment.

2. ACO Configuration

The Ant Colony Optimization algorithm was configured with 100 ants and executed over 100 iterations. The pheromone influence factor (α) was set to 1, the evaporation rate (ρ) to 0.15, and the pheromone update constant (Q) to 1. For each model, the hyperparameter search space was discretized into 20 bins per parameter to enable probabilistic selection guided by pheromone intensity.

3. Fitness Function Design

A model-specific fitness function was constructed following a common structure. For each candidate hyperparameter set proposed by an ant:

1. The corresponding model was trained on the training dataset.
2. The coefficient of determination (R^2) was calculated on the validation dataset.
3. A mild complexity penalty was applied to discourage overfitting.
4. If model training failed, a large negative fitness value was assigned to ensure algorithmic stability.

4. Optimization Process

During each ACO iteration, ants probabilistically selected hyperparameter combinations based on current pheromone distributions. After fitness evaluation, global pheromone evaporation was applied to all trails. The best-performing ant reinforced its selected path by depositing additional pheromone proportional to its fitness score. The global best solution was continuously updated and retained throughout the optimization process.

5. Final Model Training and Evaluation

After convergence of the ACO algorithm, the optimal hyperparameters were extracted. A final model was then trained using the combined training and validation datasets. Model performance was evaluated on training, validation, and test sets to assess predictive accuracy and generalization capability.

6. Visualization and Convergence Analysis

For each model, actual versus predicted CO₂ solubility values were plotted for all data subsets, including a parity line ($y = x$) to visually assess prediction accuracy. Additionally,

the progression of the best R^2 value across ACO iterations was plotted to demonstrate convergence behavior and optimization efficiency.

2.1.2. Model-Specific Hyperparameter Spaces

Although the ACO optimization procedure remained identical across models, the hyperparameter search spaces differed according to model structure:

Neural Network (NN)

The optimized hyperparameters included the number of hidden layers (1–4), number of neurons per hidden layer (5–100), and L2 regularization coefficient (10^{-5} – 10^{-1}). Model complexity was defined as the product of the number of layers and neurons.

Decision Tree (DT)

The optimized parameters were minimum leaf size (1–20), minimum parent size (5–40), and maximum number of splits (10–150). Complexity was assessed based on the number of tree nodes.

Gradient Boosting Machine (GBM)

The search space included number of trees (50–300), learning rate (0.01–0.30), and maximum number of splits (2–20). A mild penalty proportional to the number of trees was applied to encourage computational efficiency.

Support Vector Regression (SVR)

The optimized parameters were box constraint (C : 0.01–1000), epsilon (ϵ : 0.001–1), and kernel scale (0.001–100) using a radial basis function kernel. Feature standardization was applied prior to training. Model complexity was mildly penalized based on the C parameter.

2.2. Data Collection

In this study, 556 samples were collected to evaluate and predict CO_2 solubility in a solid–liquid equilibrium system containing inorganic cations and anions. The dataset includes 14 parameters such as temperature, pressure, and ion concentrations (carbonate, sodium, calcium, magnesium, potassium, chloride, sulfate, bicarbonate, iron, strontium, etc.). The dataset used in this study consists of experimentally measured CO_2 solubility values in aqueous mineral solutions across a wide range of temperatures (273–453 K), pressures (0.0021–100 MPa), and ionic compositions. The ionic composition of the solutions included the following species: sodium (Na^+), potassium (K^+), magnesium (Mg^{2+}), calcium (Ca^{2+}), chloride (Cl^-), sulfate (SO_4^{2-}), bicarbonate (HCO_3^-), bromide (Br^-), iron(II) (Fe^{2+}), strontium (Sr^{2+}), and ammonium (NH_4^+). To ensure data quality and consistency, a rigorous preprocessing protocol was applied. This involved standardizing all measurement units across variables and identifying and removing statistical outliers that could introduce bias into the analysis.

We split our data randomly into three groups: one to train the model, one to validate and adjust it, and one to test its final performance. This method ensures that our results are consistent and that different models can be compared fairly. The references for the data used in the training, validation, and test sets are [13–22].

3. Results and Discussion

This study implemented and compared four machine learning models—Neural Network (NN), Decision Tree (DT), Gradient Boosting Machine (GBM), and Support Vector Regression (SVR)—all optimized using the Ant Colony Optimization (ACO) algorithm, to

predict carbon dioxide solubility in mineral compounds. The Pearson correlation matrix of all input features and CO₂ solubility is presented in Figure 1. The matrix quantifies linear pairwise dependencies among ionic species, thermodynamic variables, and the target variable. A strong positive linear correlation between pressure and CO₂ solubility ($r = 0.57$) is observed, representing the highest correlation with the target variable. This confirms that pressure is the dominant linear driver of solubility within the investigated range and is consistent with the thermodynamic expectation of increased gas dissolution under elevated pressure. Temperature exhibits a moderate negative correlation with CO₂ solubility ($r = -0.29$), indicating that solubility decreases with increasing temperature. This trend aligns with the exothermic nature of gas dissolution in aqueous systems and further supports the physical consistency of the dataset. Among ionic species, HCO₃[−] shows a moderate positive correlation with CO₂ solubility ($r = 0.42$), while CO₃^{2−} also presents a positive correlation ($r = 0.35$). These values are higher than those of most other ions and suggest that carbonate system components are more directly associated with dissolved CO₂ levels, likely due to equilibrium interactions within the carbonate–bicarbonate system. In contrast, NH₄⁺ demonstrates a moderate negative correlation with solubility ($r = -0.37$). Sulfate (SO₄^{2−}) shows a weak negative relationship ($r = -0.13$), while K⁺ exhibits a weak-to-moderate positive correlation ($r = 0.29$). The remaining monovalent and divalent cations (Cl[−], Na⁺, Ca²⁺, Mg²⁺, Sr²⁺, Fe²⁺) display correlations close to zero ($|r| \approx 0.00$ – 0.04), indicating negligible linear dependence with CO₂ solubility when evaluated individually.

A notable feature of the matrix is the extremely high intercorrelation among several ionic species. For example, Cl[−], Na⁺, Ca²⁺, Mg²⁺, and Fe²⁺ show near-perfect correlations ($r \approx 0.95$ – 1.00) with one another. Similarly, Br[−] and Sr²⁺ are strongly correlated ($r = 0.97$). These strong interdependencies indicate significant multicollinearity within the ionic composition variables, likely arising from common brine formulations or charge-balanced salt systems reported in the literature sources. Such multicollinearity suggests that individual linear coefficients may not independently represent the physicochemical contribution of each ion.

Importantly, because Pearson correlation captures only linear dependence, weak pairwise correlations do not necessarily imply negligible influence. The complex electrolyte–gas interactions governing CO₂ dissolution are inherently nonlinear and multivariate. Therefore, while the correlation matrix provides useful exploratory insight, it does not fully describe the underlying dependencies, further justifying the application of nonlinear machine learning techniques in this study. In solid–liquid–gas equilibrium systems involving CO₂ dissolution in brines, interionic and ion–molecule interactions are inherently nonlinear and composition-dependent. Even if two ions are linearly correlated in concentration across the compiled dataset, their physicochemical influence on activity coefficients, complex formation, or carbonate equilibria may differ substantially.

Figure 2 illustrates the distribution plots of the 556 collected samples, offering important insight into the statistical structure and thermodynamic consistency of the dataset. The distribution plots of the 556 collected samples provide important insight into the statistical structure and thermodynamic consistency of the dataset. The ionic composition variables exhibit clustered and discrete concentration levels, indicating that the dataset was compiled from multiple literature sources with distinct brine formulations rather than from a single continuous experimental campaign. For several ions, including NH₄⁺, Cl[−], Na⁺, Ca²⁺, and Mg²⁺, a considerable number of samples are concentrated at or near zero concentration. This suggests that many reported brine systems did not contain these species, while other subsets of the dataset represent specific chemical environments with elevated concentrations. In contrast, ions such as SO₄^{2−}, HCO₃[−], and CO₃^{2−} display multiple clustered concentration levels, reflecting variations in carbonate–sulfate equilibria among different

brine systems. Trace species such as Fe^{2+} and Br^- exhibit narrow concentration ranges, indicating limited variability across the compiled studies.

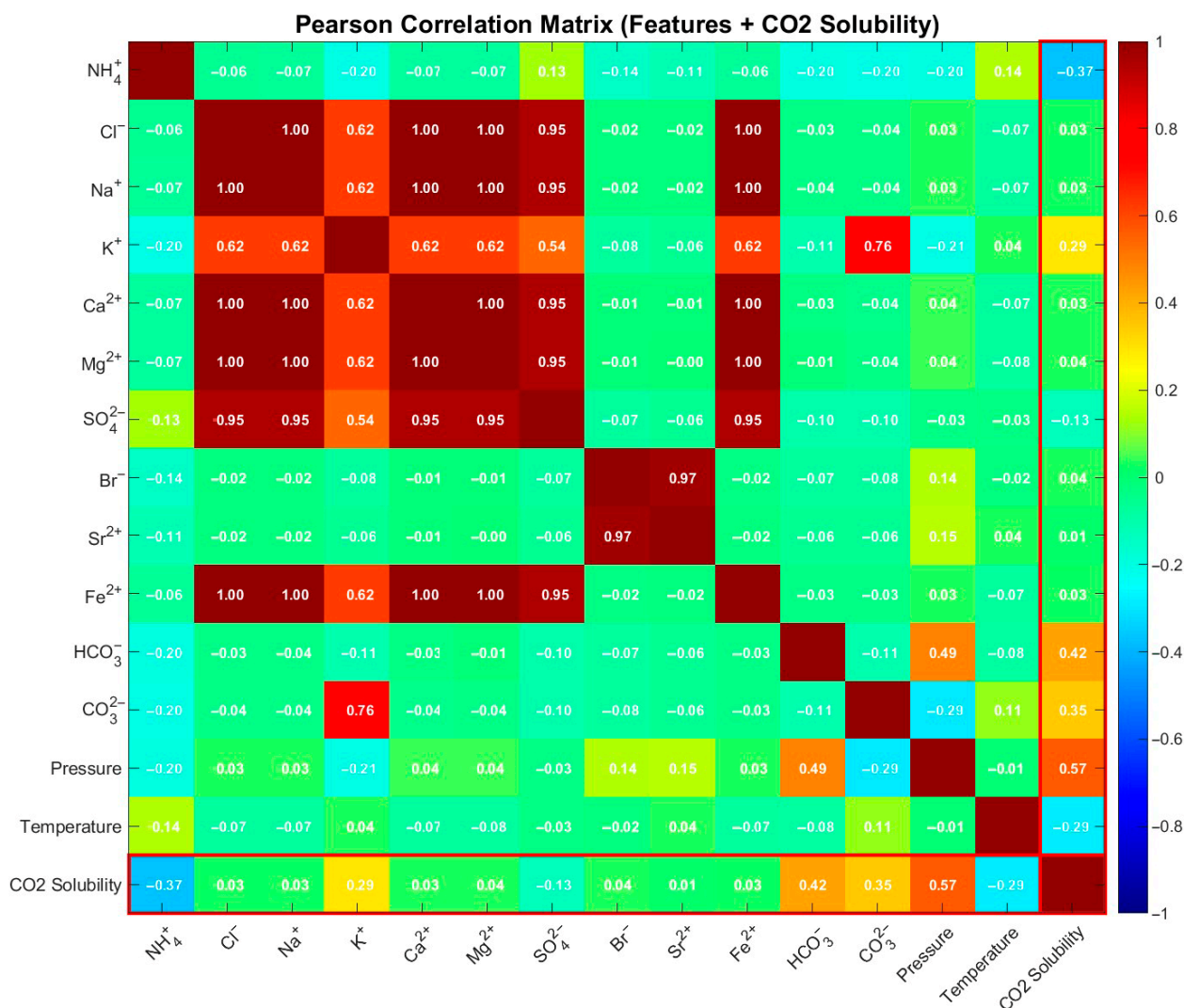


Figure 1. The Pearson correlation matrix of all input features and CO₂ solubility.

The pressure–solubility relationship demonstrates a clear positive trend. CO₂ solubility increases systematically with increasing pressure across the investigated range (approximately 0–45 MPa). The relationship appears approximately linear at lower pressures and gradually transitions to a milder slope at higher pressures, which is consistent with the expected behavior of gas dissolution under increasing compressibility effects. This confirms the thermodynamic reliability of the collected data and aligns with Henry’s law behavior within the investigated range. In contrast, the temperature–solubility plot reveals an overall negative dependence of CO₂ solubility on temperature across the range of approximately 280–450 K. Higher temperatures correspond to reduced solubility levels, which is consistent with the exothermic nature of gas dissolution in aqueous systems. The data coverage across this wide temperature interval ensures that the developed model is applicable to a broad spectrum of subsurface and CCUS-related conditions.

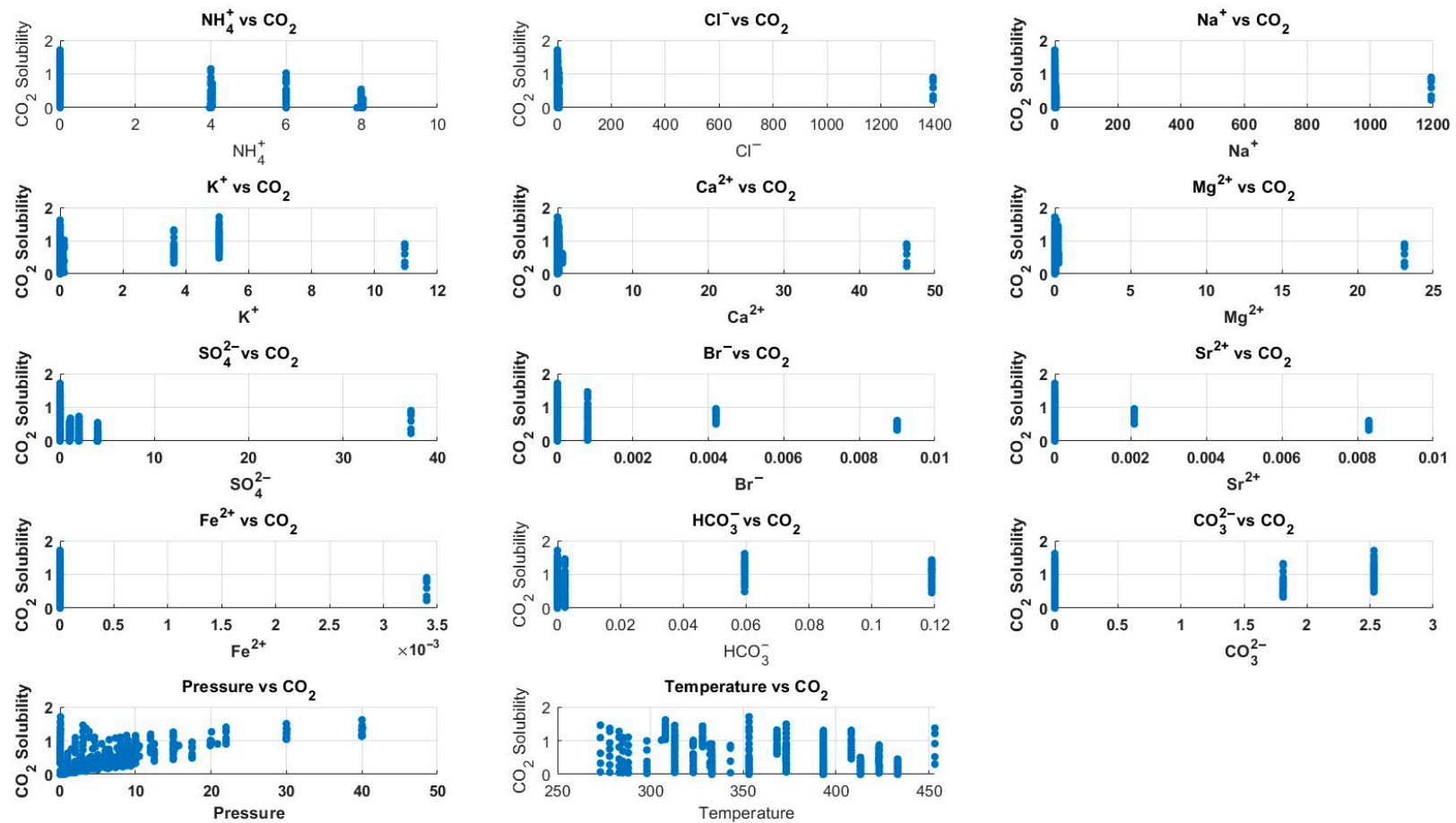


Figure 2. Effect of ionic composition, pressure and temperature on CO_2 solubility.

The scatter plots of individual ionic species versus CO₂ solubility indicate predominantly nonlinear and composition-dependent behavior. Divalent cations such as Ca²⁺, Mg²⁺, and Sr²⁺ tend to correspond to slightly reduced solubility levels at higher concentrations, suggesting a salting-out effect driven by increased ionic strength. Carbonate and bicarbonate species (HCO₃[−] and CO₃^{2−}) display structured clusters, reflecting their involvement in chemical equilibrium reactions that influence dissolved CO₂ speciation. Overall, the absence of simple linear trends between most individual ions and solubility suggests that multivariate nonlinear modeling approaches, such as machine learning, are appropriate for capturing the complex interactions within the system. Importantly, despite the clustered nature of several ionic variables, the dataset spans a broad operational domain in pressure, temperature, and chemical composition. This diversity supports the robustness and generalization capability of the predictive model developed in this study.

In Figure 3, the predicted results of carbon dioxide solubility in the presence of minerals are presented based on training, validation, and testing, using the Neural Network Algorithm optimized by Ant Colony Optimization. The following optimal parameters were obtained for the neural network:

Number of neurons: 15;

Number of layers: 3;

L2 Regularization: 1.0000×10^{-5} .

The model performance, expressed as R² scores, is as follows:

Training R²: 0.968 (RMSE = 0.0701, MAE = 0.0268);

Validation R²: 0.887 (RMSE = 0.1200, MAE = 0.0473);

Testing R²: 0.930 (RMSE = 0.1083, MAE = 0.0572).

These results indicate that the optimized neural network demonstrates strong predictive ability and good generalization across training, validation, and test datasets.

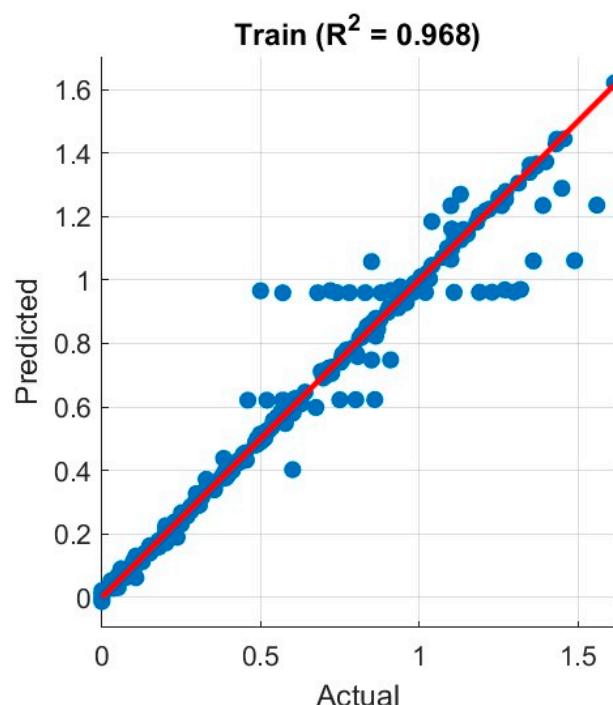


Figure 3. Cont.

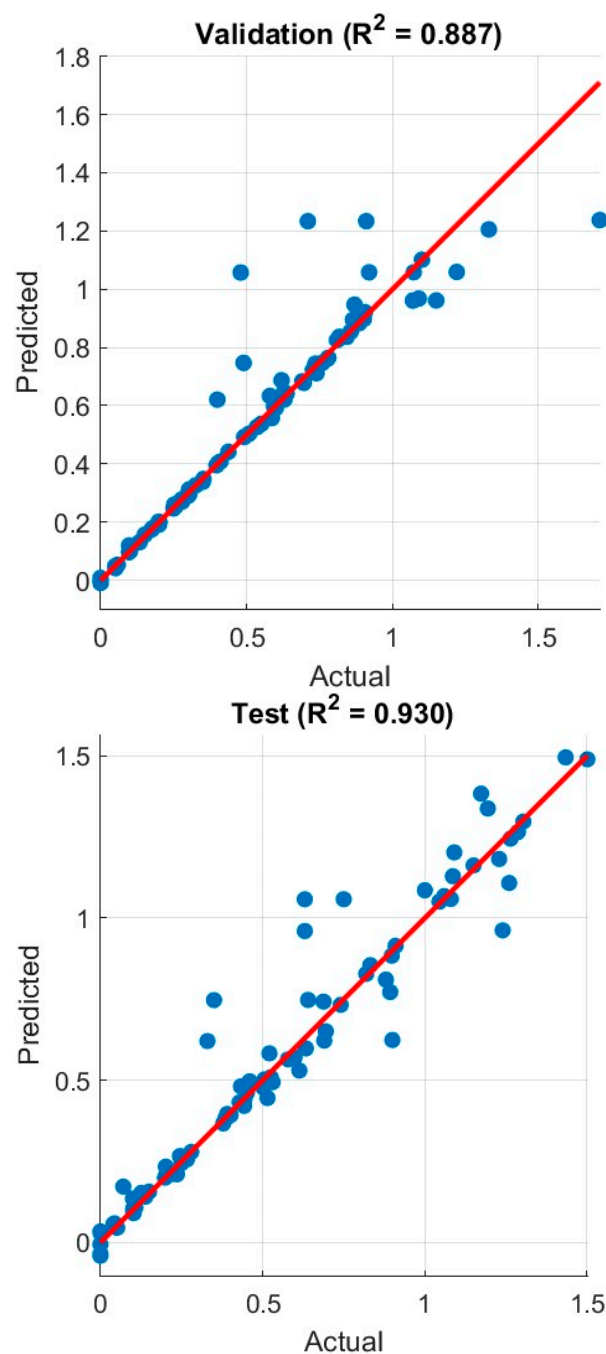


Figure 3. Optimized Neural Network Regression Model: Actual vs. Predicted CO₂ Solubility for Training, Validation, and Test Sets.

In Figure 4, the prediction results for carbon dioxide solubility in mineral compounds are presented based on training, validation, and testing data, utilizing a Decision Tree model optimized by Ant Colony Optimization (ACO). The model was fine-tuned with the following optimal parameters: MinLeaf of 4, MinParent of 23, and MaxSplit of 135. The resulting decision tree contains 67 total nodes, with 34 leaf nodes, and reaches a maximum depth of 7. The performance evaluation using the R^2 metric demonstrates the model's effectiveness across different datasets:

Training R^2 : 0.914 (RMSE = 0.1152, MAE = 0.0812);

Validation R^2 : 0.848 (RMSE = 0.1390, MAE = 0.0920);

Testing R^2 : 0.912 (RMSE = 0.1215, MAE = 0.0893).

These results confirm that the ACO-optimized decision tree is accurate and reliable for predicting CO₂ solubility in mineral systems, with a balanced structure that mitigates overfitting while maintaining explanatory power.

In Figure 5, the predicted solubility of carbon dioxide in mineral compounds based on training, validation, and testing data is presented using a Support Vector Regression (SVR) model optimized by Ant Colony Optimization (ACO).

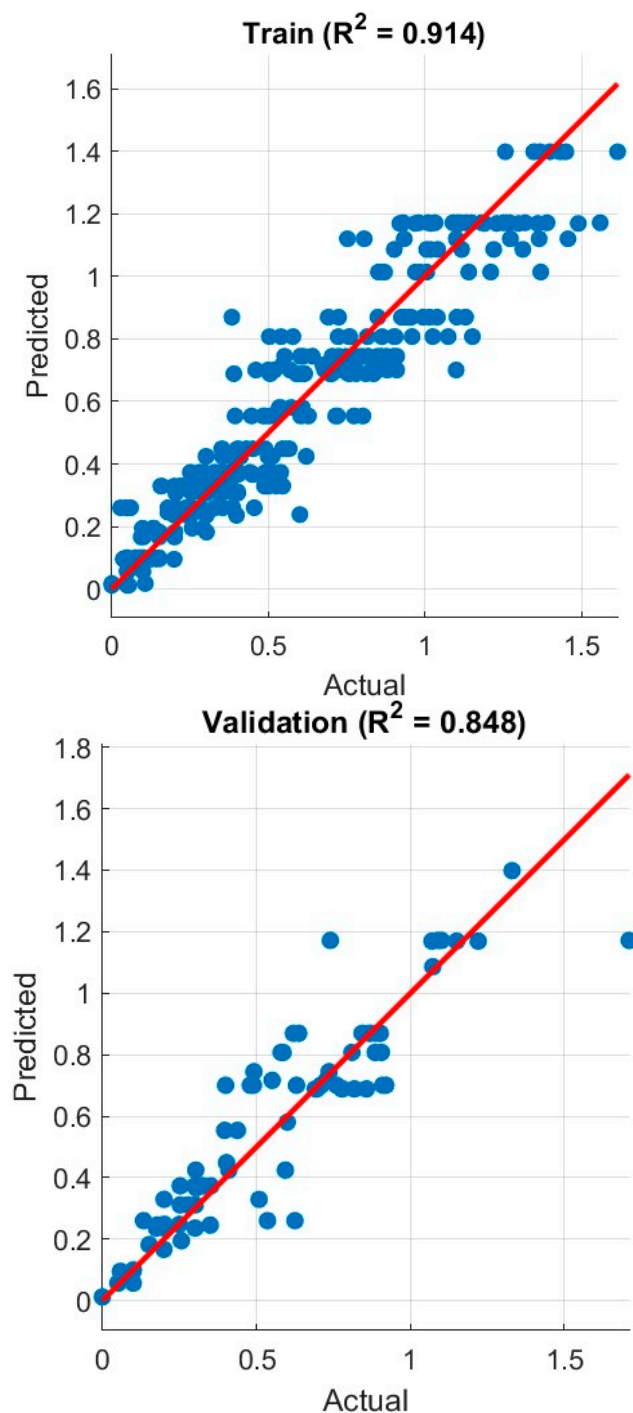


Figure 4. Cont.

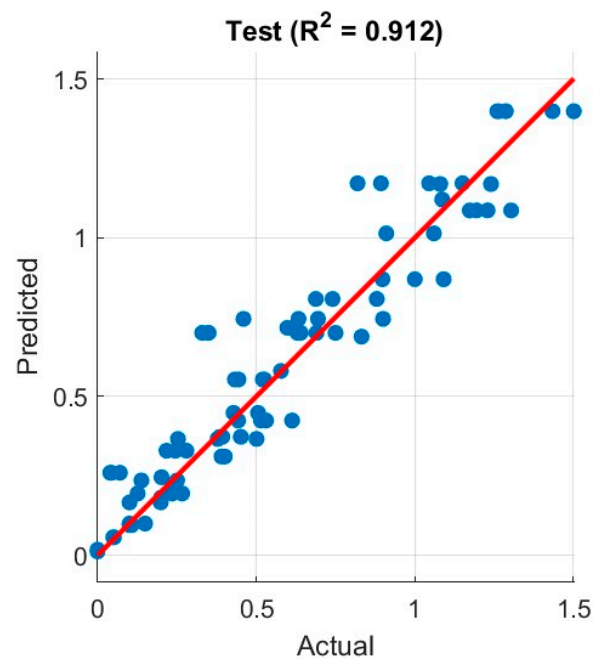


Figure 4. Prediction Performance of the ACO-Optimized Decision Tree for CO₂ Solubility in Mineral Composites: Training, Validation, and Test Sets.

The following optimal hyperparameters were identified for the SVR model:

C (Regularization Parameter): 52.6411;

Epsilon (ϵ): 0.001;

Gamma (γ): 5.2641.

The model's performance was evaluated using the R^2 metric, yielding the following results:

Training R^2 : 0.942 (RMSE = 0.0948, MAE = 0.0432);

Validation R^2 : 0.854 (RMSE = 0.1363, MAE = 0.0659);

Testing R^2 : 0.889 (RMSE = 0.1363, MAE = 0.0659).

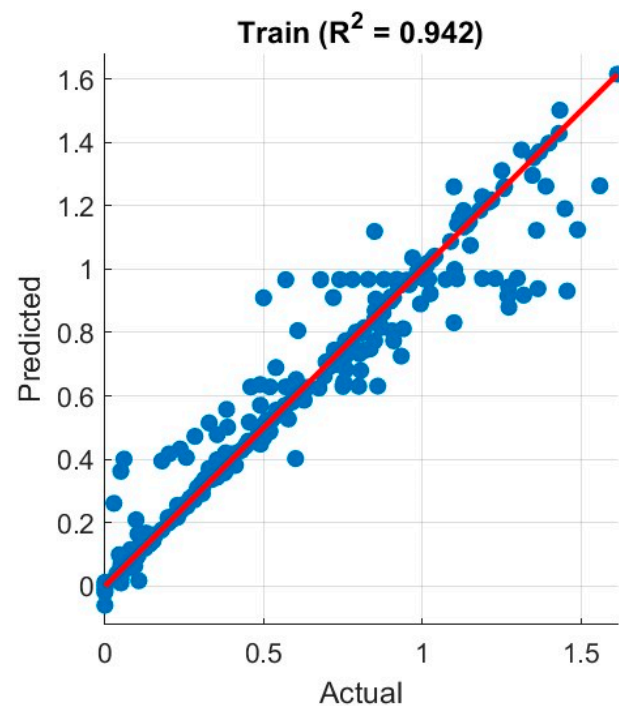


Figure 5. Cont.

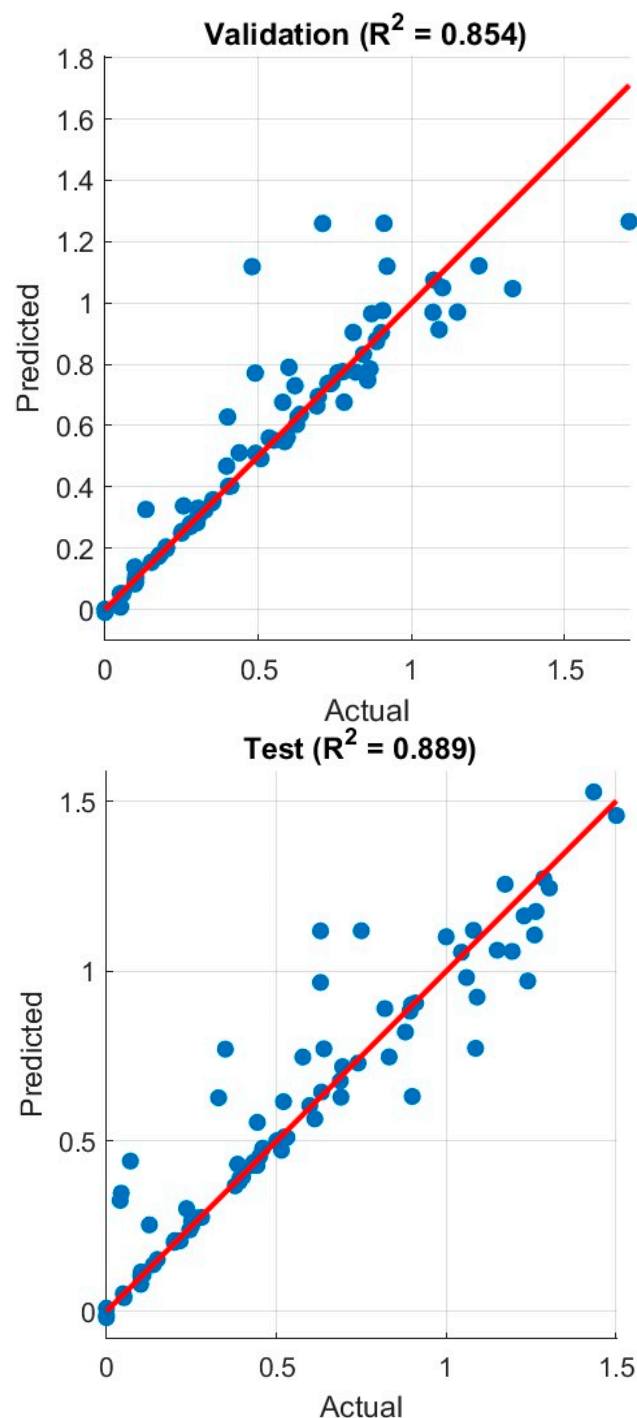


Figure 5. Accuracy of Support Vector Regression (SVR) Optimized by ACO for Forecasting CO₂ Solubility: Training, Validation, and Test Sets.

These results indicate that the ACO-optimized SVR model successfully learned complex relationships within the training data while maintaining a strong ability to generalize. The high training score reflects an excellent fit, and the competitive testing score demonstrates the model's robustness and reliability for predicting CO₂ solubility in new, unseen mineral compositions.

In Figure 6, the prediction results for carbon dioxide solubility in mineral compounds are presented based on training, validation, and testing, employing a Gradient Boosting Machine (GBM) model optimized by Ant Colony Optimization (ACO). The model was fine-tuned with the following optimal parameters: MinLeaf of 4, MinParent of 23,

MaxSplit of 13, nTrees of 50, and a Learning Rate of 0.2695. The model demonstrated exceptional performance:

Training R^2 : 0.995 (RMSE = 0.0285, MAE = 0.0178);

Validation R^2 : 0.995 (RMSE = 0.0261, MAE = 0.0193);

Testing R^2 : 0.986 (RMSE = 0.0478, MAE = 0.0362).

These results underscore the GBM model's superior capability in capturing the complex relationships governing CO_2 solubility in mineral systems, establishing it as a highly reliable tool for this application.

All models achieved strong predictive performance, with test R^2 scores ranging from 0.889 to 0.986. The Gradient Boosting Machine emerged as the top performer, achieving near-perfect scores on both training and validation sets ($R^2 = 0.995$) and an exceptional test score of 0.986. This indicates outstanding generalization capability with minimal overfitting. Its primary strength lies in its exceptional accuracy and stability. The ensemble approach of sequentially correcting errors from multiple shallow trees, combined with an optimal learning rate (0.2695) and tree count (50), allowed it to capture complex, non-linear relationships in the data without overfitting. The Neural Network also performed well (Test $R^2 = 0.930$), showing strong improvement with its deeper architecture (3 layers, 15 neurons) compared to simpler networks. The model demonstrated balanced performance across all datasets with relatively low error rates. The Decision Tree performed excellently on the test set ($R^2 = 0.912$), showing consistent results between training and testing. The key strength of the Decision Tree is its interpretability and transparency. With 67 nodes and a depth of 7, the model's structure and decision paths can be visualized and understood, providing insight into which mineral features most influence solubility. The Support Vector Regression delivered solid results (Test $R^2 = 0.889$), demonstrating reliable robustness. This model's core strength is its generalization reliability. The optimal parameters ($C = 52.64$, $\epsilon = 0.001$, $\gamma = 5.26$) allowed the RBF kernel to model complex relationships, while the regularization parameter prevented overfitting.

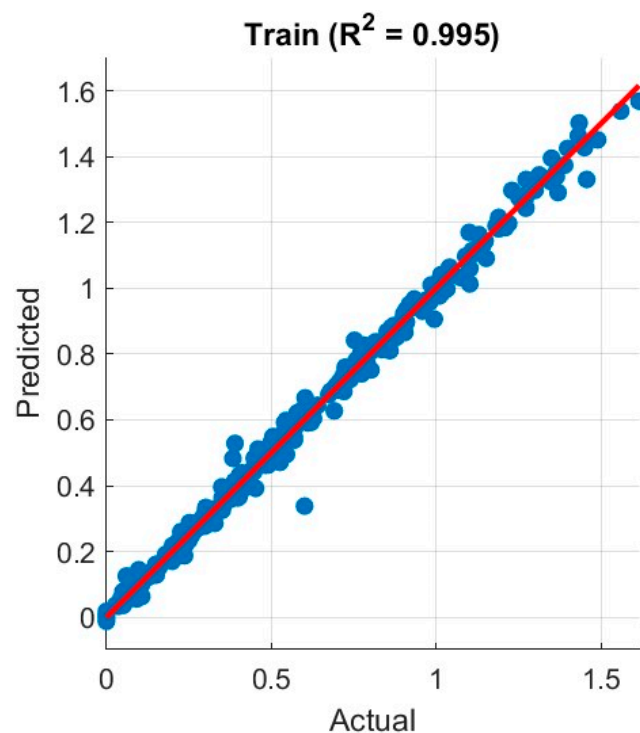


Figure 6. Cont.

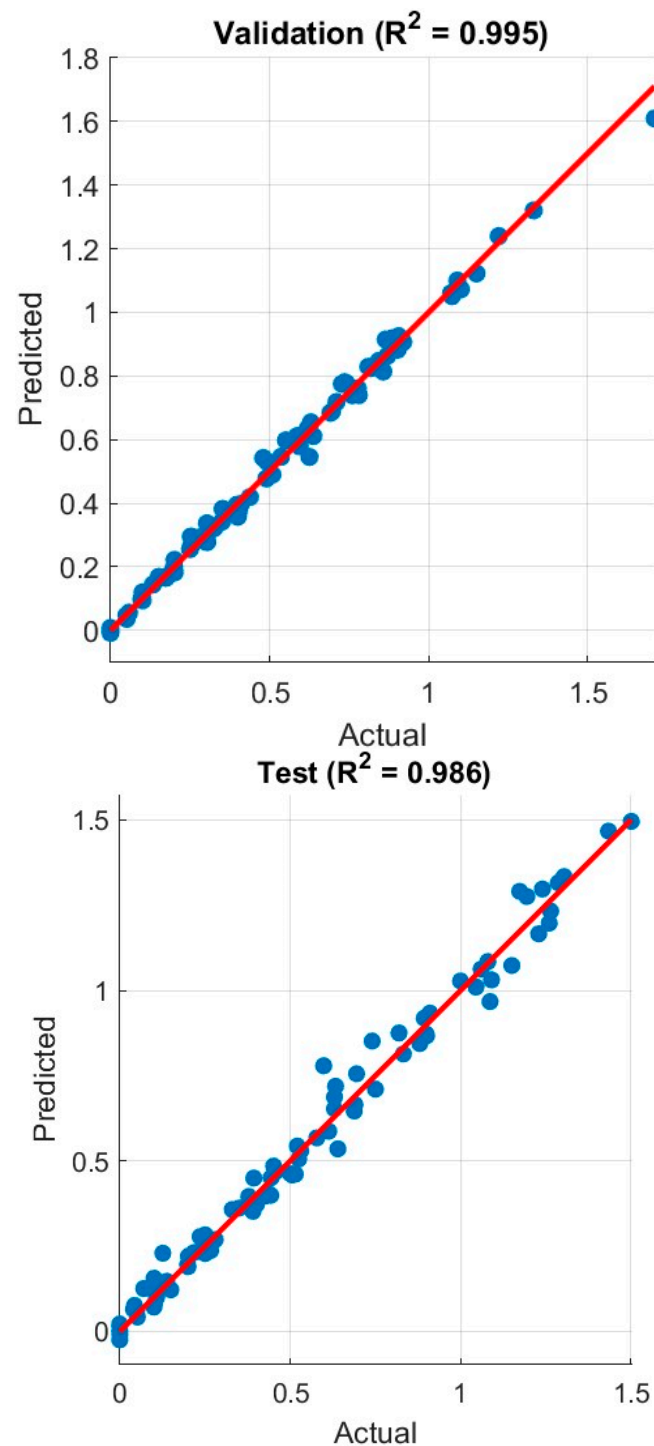


Figure 6. Optimized Gradient Boosting Machine Model: Actual vs. Predicted CO₂ Solubility for Training, Validation, and Test Sets.

The choice of the optimal model depends on the project's specific priority. If maximum predictive accuracy is paramount, the GBM model is unequivocally the best choice. If model interpretability and transparency are critical for scientific understanding, the Decision Tree is highly valuable. For scenarios requiring a robust neural network with strong performance, the optimized Neural Network with 3 layers is an excellent candidate. The SVR offers balanced performance when consistent generalization across all datasets is needed. This comparative analysis demonstrates that ACO is an effective metaheuristic

for tuning diverse ML models, successfully navigating different hyperparameter spaces to enhance their predictive performance for a complex scientific problem.

In Table 1, a comparison of model performance with and without ACO optimization is presented. The results clearly show that using Ant Colony Optimization to tune the machine learning models consistently improved their performance on unseen test data. This means the optimization algorithm successfully found better hyperparameters for each model, making them more reliable for predicting CO₂ solubility in new situations.

Table 1. Comparison of Model Performance With and Without ACO Optimization.

Model	Dataset	Without ACO (R ²)	With ACO (R ²)
ANN	Training	0.965	0.968
	Validation	0.835	0.887
	Test	0.901	0.930
DT	Training	0.935	0.914
	Validation	0.832	0.848
	Test	0.897	0.912
SVR	Training	0.967	0.942
	Validation	0.803	0.854
	Test	0.846	0.889
GBM	Training	0.988	0.995
	Validation	0.957	0.995
	Test	0.979	0.986

For the Neural Network, ACO had a significant impact. While the training score stayed nearly the same, the test R² jumped from 0.901 to 0.930. Even more importantly, the validation score improved dramatically from 0.835 to 0.887. This shows that the optimized network, with its three layers and fifteen neurons, generalized much better and was no longer overfitting to the training data. The Support Vector Regression model saw the biggest benefit from ACO. Its test score increased substantially from 0.846 to 0.889, and the validation score rose from 0.803 to 0.854. This large improvement indicates that ACO was essential for finding the right C, epsilon, and gamma values, which unlocked the SVR model's true potential and made it far more robust. The Decision Tree model also became more accurate and stable after optimization. Its test score increased from 0.897 to 0.912. The Gradient Boosting Machine was already the best performer, but ACO still managed to improve it. The test score increased from an already excellent 0.979 to a near-perfect 0.986. The most notable change was in the validation score, which jumped from 0.957 to match the training score at 0.995. This shows that ACO perfectly balanced the model's learning rate and tree count, eliminating any small amount of overfitting and creating an exceptionally stable and accurate model.

4. Conclusions

This study presented a unified hybrid framework integrating Ant Colony Optimization with four machine learning models, Neural Network, Decision Tree, Support Vector Regression, and Gradient Boosting Machine, for systematic hyperparameter tuning in CO₂ solubility prediction. The consistent application of ACO across all models enabled a fair comparative evaluation under identical optimization conditions. The results demonstrate that ACO effectively enhances model performance by efficiently exploring complex hyperparameter spaces. Among the evaluated algorithms, GBM achieved the highest predictive accuracy with a test R² of 0.986, confirming its superior capability in modeling nonlinear solubility behavior. The Neural Network also performed strongly, achieving a test R² of 0.930 with its optimized three-layer architecture. The Decision Tree model provided

competitive performance with a test R^2 of 0.912 and clear interpretability advantages, while SVR showed reliable generalization with a test R^2 of 0.889. The primary novelty of this work lies in implementing a unified ACO-driven optimization strategy consistently across multiple machine learning architectures, reducing manual tuning bias and improving predictive robustness. The findings indicate that ensemble-based models, particularly GBM, are the most suitable for high-accuracy CO₂ solubility prediction when maximum precision is required. However, the strong neural network performance demonstrates that deeper architectures can also achieve excellent results when properly optimized. Interpretable models like Decision Tree remain valuable when transparency is needed. This framework offers a transferable and computationally efficient approach for predictive modeling in geochemical and subsurface engineering applications.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/w18060662/s1>, Table S1: Data.

Author Contributions: Conceptualization, S.H.H. and F.T.; methodology, S.H.H., F.T. and S.P.; software, S.H.H., F.T. and S.P.; validation, S.H.H.; formal analysis, S.H.H. and F.T.; investigation, S.H.H. and F.T.; writing—original draft preparation, S.H.H., F.T. and S.P.; writing—review and editing, S.H.H., F.T. and S.P.; supervision, F.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

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

References

1. Luo, A.; Li, Y.; Chen, X.; Zhu, Z.; Peng, Y. Review of CO₂ sequestration mechanism in saline aquifers. *Nat. Gas Ind. B* **2022**, *9*, 383–393. [\[CrossRef\]](#)
2. Izadpanahi, A.; Blunt, M.J.; Kumar, N.; Ali, M.; Tassinari, C.C.G.; Sampaio, M.A. A review of carbon storage in saline aquifers: Mechanisms, prerequisites, and key considerations. *Fuel* **2024**, *369*, 131744. [\[CrossRef\]](#)
3. Youssef, M.; Isah, M.; Rezk, M.; Gbadamosi, A.; Raza, A.; Mahmoud, M. Unlocking the Potential of CO₂ Storage in Saline Aquifers: Challenges, Knowledge Gaps, and Future Directions for Large-Scale Storage. *Carbon Capture Sci. Technol.* **2025**, *16*, 100460. [\[CrossRef\]](#)
4. Emami-Meybodi, H.; Hassanzadeh, H. Two-phase convective mixing under a buoyant plume of CO₂ in deep saline aquifers. *Adv. Water Resour.* **2015**, *76*, 55–71. [\[CrossRef\]](#)
5. Zhang, D.; Song, J. Mechanisms for geological carbon sequestration. *Procedia IUTAM* **2014**, *10*, 319–327. [\[CrossRef\]](#)
6. Hashemi, S.H.; Torabi, F.; Tontiwachwuthikul, P. Machine Learning-Driven Prediction of CO₂ Solubility in Brine: A Hybrid Grey Wolf Optimizer (GWO)-Assisted Gaussian Process Regression (GPR) Approach. *Energies* **2025**, *18*, 4205. [\[CrossRef\]](#)
7. Davoodi, S.; Longe, P.O.; Thanh, H.V.; Mehrad, M.; Mohammadi, A.H.; Burnaev, E. Machine-learning models for predicting CO₂ solubility in various brine systems: Implications for carbon geo-storage. *J. Mol. Liq.* **2025**, *435*, 128122. [\[CrossRef\]](#)
8. Wei, W.; Lu, P.; Zhu, C.; Luo, P.; Mesdour, R. Advanced Machine Learning Models for CO₂ and H₂S Solubility in Water and NaCl Brine: Implications for Geoenergy Extraction and Carbon Storage. *Energy Fuels* **2024**, *38*, 11119–11136. [\[CrossRef\]](#)
9. Yang, S.; Wang, D.; Dong, Z.; Li, Y.; Du, D. ANN prediction of the CO₂ solubility in water and brine under reservoir conditions. *AIMS Geosci.* **2025**, *11*, 201–227. [\[CrossRef\]](#)
10. Bhattacharjee, R.; Botchway, K.; Pashin, J.; Chakraborty, G.; Bikkina, P. Machine learning-based prediction of CO₂ fugacity coefficients: Application to estimation of CO₂ solubility in aqueous brines as a function of pressure, temperature, and salinity. *Int. J. Greenh. Gas Control* **2023**, *128*, 103971. [\[CrossRef\]](#)
11. Sadeghi, A.; Salami, H.; Taghikhani, V.; Robert, M. A comprehensive study on CO₂ solubility in brine: Thermodynamic-based and neural network modeling. *Fluid Phase Equilibria* **2015**, *403*, 153–159. [\[CrossRef\]](#)
12. Zou, X.; Zhu, Y.; Lv, J.; Zhou, Y.; Ding, B.; Liu, W.; Xiao, K.; Zhang, Q. Toward Estimating CO₂ Solubility in Pure Water and Brine Using Cascade Forward Neural Network and Generalized Regression Neural Network: Application to CO₂ Dissolution Trapping in Saline Aquifers. *ACS Omega* **2024**, *9*, 4705–4720. [\[CrossRef\]](#) [\[PubMed\]](#)

13. Rumpf, B.; Nicolaisen, H.; Maurer, G. Solubility of carbon dioxide in aqueous solutions of ammonium chloride at temperatures from 313 K to 433 K and pressures up to 10 MPa. *Berichte Bunsenges. Phys. Chem.* **1994**, *98*, 1077–1081. [[CrossRef](#)]
14. El-Maghraby, R.; Pentland, C.; Iglauer, S.; Blunt, M. A fast method to equilibrate carbon dioxide with brine at high pressure and elevated temperature including solubility measurements. *J. Supercrit. Fluids* **2012**, *62*, 55–59. [[CrossRef](#)]
15. Zhao, H.; Dillmore, R.; Allen, D.E.; Hedges, S.W.; Soong, Y.; Lvov, S.N. Measurement and modeling of CO₂ solubility in natural and synthetic formation brines for CO₂ sequestration. *Environ. Sci. Technol.* **2015**, *49*, 1972–1980. [[CrossRef](#)]
16. Li, Z.; Dong, M.; Li, S.; Dai, L. Densities and Solubilities for Binary Systems of Carbon Dioxide + Water and Carbon Dioxide + Brine at 59 °C and Pressures to 29 MPa. *J. Chem. Eng. Data* **2004**, *49*, 1026–1031. [[CrossRef](#)]
17. Poulain, M.; Messabeb, H.; Lach, A.; Contamine, F.; Cézac, P.; Serin, J.-P.; Dupin, J.-C.; Martinez, H. Experimental Measurements of Carbon Dioxide Solubility in Na–Ca–K–Cl Solutions at High Temperatures and Pressures up to 20 MPa. *J. Chem. Eng. Data* **2019**, *64*, 2497–2503. [[CrossRef](#)]
18. Rumpf, B.; Maurer, G. An Experimental and Theoretical Investigation on the Solubility of Carbon Dioxide in Aqueous Solutions of Strong Electrolytes. *Berichte Bunsenges. Phys. Chem.* **1993**, *97*, 85–97. [[CrossRef](#)]
19. Cruz, J.L.; Neyrolles, E.; Contamine, F.; Cézac, P. Experimental Study of Carbon Dioxide Solubility in Sodium Chloride and Calcium Chloride Brines at 333.15 and 453.15 K for Pressures up to 40 MPa. *J. Chem. Eng. Data* **2021**, *66*, 249–261. [[CrossRef](#)]
20. Stewart, P.B.; Munjal, P. Solubility of Carbon Dioxide in Pure Water, Synthetic Sea Water, and Synthetic Sea Water Concentrates at –50 to 250 C. and 10- to 45-Atm. Pressure. *J. Chem. Eng. Data* **1970**, *15*, 67–71. [[CrossRef](#)]
21. Tang, Y.; Bian, X.; Du, Z.; Wang, C. Measurement and prediction model of carbon dioxide solubility in aqueous solutions containing bicarbonate anion. *Fluid Phase Equilibria* **2015**, *386*, 56–64. [[CrossRef](#)]
22. Grimekis, D.; Giannoulidis, S.; Manou, K.; Panopoulos, K.; Karellas, S. Experimental investigation of CO₂ solubility and its absorption rate into promoted aqueous potassium carbonate solutions at elevated temperatures. *Int. J. Greenh. Gas Control* **2019**, *81*, 83–92. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.