

Article

Data-Driven Prediction of Caprock Breakthrough Pressure and Permeability for CO₂ Storage Integrity: A Meta-Analysis

R. G. C. Tharuksha and K. H. S. M. Sampath * 

Department of Civil Engineering, University of Moratuwa, Moratuwa 10400, Sri Lanka

* Correspondence: sampathkh@uom.lk

Abstract

Carbon capture and storage (CCS) is a key strategy for achieving net-zero emissions, with long-term storage security depending on caprock integrity. Caprocks act as sealing barriers preventing upward CO₂ migration; however, breakthrough pressure and permeability remain difficult to predict across variable geological settings. This study presents a data-driven meta-analysis of caprock sealing performance using published datasets covering confining pressure, temperature, depth, porosity, permeability, and breakthrough pressure. Missing values were treated using Multiple Imputation by Chained Equations (MICE), and predictive performance was evaluated using ordinary least squares regression, polynomial regression, neural networks, physics-informed residual/hybrid models, and least absolute shrinkage and selection operator (LASSO) regression with polynomial feature expansion. Severe multicollinearity was identified between temperature and depth, and conventional regression showed weak permeability prediction due to multicollinearity and geological heterogeneity, whereas the physics-informed residual model improved performance from $R^2 = 0.28$ to $R^2 = 0.77$ by incorporating depth- and temperature-dependent residual corrections. For breakthrough pressure, the physics-informed hybrid model showed strong fitting performance, while the LASSO-derived empirical equation provided a more generalizable prediction with a validated test R^2 of 0.71 and a mean absolute error of 2.31 MPa. The proposed framework supports preliminary screening for safe CO₂ injection pressure design and long-term storage integrity assessment.

Keywords: CO₂ sequestration; caprock integrity; meta-analysis; permeability; breakthrough pressure; physics-informed modeling



Academic Editor: Alberto Maria Gambelli

Received: 15 June 2026

Revised: 10 July 2026

Accepted: 13 July 2026

Published: 16 July 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

The increasing concentration of atmospheric carbon dioxide (CO₂) resulting from fossil fuel combustion, industrial activities, and energy-intensive development has become one of the principal drivers of global climate change. Reducing anthropogenic CO₂ emissions is therefore essential for achieving long-term climate targets and net-zero emission pathways. Among available mitigation technologies, carbon capture and storage (CCS) has received considerable attention because it provides a technically feasible and scalable approach to reducing emissions from large stationary sources, including power generation, cement production, steel manufacturing, and other energy-intensive industries. In geological CO₂ sequestration, captured CO₂ is compressed and injected into deep subsurface formations, such as saline aquifers, depleted oil and gas reservoirs, and unmineable coal seams, where it is expected to remain securely stored over geological timescales [1].

The long-term safety and effectiveness of geological CO₂ storage depend fundamentally on the integrity of the overlying caprock. Caprock formations act as low-permeability sealing barriers that prevent the upward migration of injected CO₂ from the storage reservoir into shallower formations or the atmosphere. These sealing units are commonly composed of fine-grained formations such as shale, mudstone, claystone, or evaporites. Their containment capacity is governed by a combination of petrophysical, geological, geochemical, and geo-mechanical properties. During CO₂ injection, increasing pore pressure may alter the in situ stress state, capillary sealing behavior, and fluid–rock interactions within the storage system. Consequently, caprock integrity should not be treated only as a static geological property, but as a coupled response to changing pressure, temperature, stress, and pore-structure conditions.

Permeability and breakthrough pressure are two of the most important parameters used to evaluate caprock sealing performance. Permeability describes the ability of fluids to flow through the rock matrix and is commonly used to assess potential leakage through connected pore networks, fractures, or defects. Lower permeability generally indicates better sealing capacity. However, permeability is highly sensitive to geological heterogeneity, pore structure, clay content, bedding, fractures, and effective stress conditions, making it difficult to generalize across different caprock formations. Breakthrough pressure, in contrast, represents the minimum pressure required for a non-wetting fluid such as CO₂ to overcome capillary resistance and penetrate the water-saturated pore network of the caprock. It is therefore a more direct measure of capillary sealing capacity and is particularly important for defining safe injection pressure limits in geological CO₂ storage projects [2].

Several geological and mechanical variables influence permeability and breakthrough pressure. Porosity affects pore volume and pore connectivity, thereby controlling both fluid flow potential and capillary resistance. Depth influences in situ stress conditions and is closely related to confining pressure and formation temperature. Temperature affects CO₂ density, brine properties, interfacial tension, and possible fluid–rock reactions, while confining pressure controls pore-throat closure and effective stress-dependent changes in sealing behavior. These variables are strongly interrelated, and their combined influence on caprock sealing performance may be nonlinear. As a result, conventional single-parameter correlations are often insufficient for predicting caprock behavior under realistic reservoir conditions, and multi-variable analytical frameworks are required to capture the coupled nature of these interactions.

Field-scale CCS projects have demonstrated both the feasibility and the geo-mechanical sensitivity of geological CO₂ storage. Projects such as Sleipner and Weyburn have demonstrated the feasibility of secure storage within relatively undeformed, matrix-dominated caprocks [3,4]. However, observations from the In Salah project, where injection-induced pressure reactivated pre-existing fault zones and fracture networks, highlight the limitations of matrix-only sealing assumptions [5]. These field experiences emphasize that reliable caprock integrity assessment requires predictive frameworks that ultimately account for local tectonic and diagenetic discontinuities, in addition to coupled pressure, temperature, stress, and pore-structure effects.

Despite extensive research on caprock behavior, several limitations remain in the existing literature. Many studies focus on isolated relationships between individual parameters, such as porosity–permeability relationships or the influence of confining pressure on breakthrough pressure. Although such investigations provide valuable insights, their findings are often formation-specific and difficult to generalize to different geological settings. In addition, experimental datasets are frequently incomplete because published studies do not consistently report the same variables. Some studies provide permeability and porosity data but omit breakthrough pressure, confining pressure, or temperature, while others

report breakthrough pressure without sufficient information on depth or pore-structure characteristics. This fragmented reporting limits the development of robust predictive models and reduces confidence in applying empirical correlations to new storage sites [6,7].

Missing and inconsistent data represent an additional challenge in caprock integrity research. Conventional approaches, such as complete-case deletion, reduce the available sample size and may introduce selection bias, particularly when datasets are already limited. Simple imputation methods, such as mean substitution, may also distort the statistical relationships among variables. Multiple Imputation by Chained Equations (MICE) provides a more rigorous approach by estimating missing values based on the conditional relationships among all available variables [8]. This makes MICE particularly suitable for meta-analysis studies that integrate heterogeneous data from multiple published sources while attempting to preserve inter-variable dependencies.

Recent advances in data-driven modeling offer new opportunities for improving caprock integrity assessment. Ordinary least squares regression and polynomial regression provide interpretable baseline relationships, but their performance may be limited when predictor variables are strongly correlated or when the underlying behavior is nonlinear. Neural networks can capture complex patterns, but their generalization ability is often constrained when datasets are small, as is common in geo-mechanical and petrophysical studies. Physics-informed and hybrid residual models provide a useful alternative by combining domain knowledge with data-driven correction terms, thereby improving predictive accuracy while retaining a stronger connection to the physical behavior of caprock systems. Regularized regression methods, such as Least Absolute Shrinkage and Selection Operator (LASSO) regression, can further improve model stability, reduce overfitting, and support the development of interpretable empirical equations for engineering use [9].

Although previous studies have examined permeability and breakthrough pressure separately, there remains a clear need for an integrated data-driven study that evaluates both sealing indicators within a common analytical framework. Permeability reflects the potential for fluid flow through the caprock matrix, whereas breakthrough pressure defines the threshold pressure required for CO₂ entry. Assessing both parameters together allows a more complete characterization of caprock sealing performance and provides a stronger basis for preliminary site screening and safe injection pressure design.

To address these limitations, this study presents an integrated data-driven meta-analysis of caprock permeability and breakthrough pressure using datasets compiled from the published literature. Porosity, depth, temperature, and confining pressure are considered as the primary governing variables, while missing values are treated using MICE to improve dataset completeness and preserve inter-variable relationships. A progressive modeling framework is applied, including ordinary least squares regression, polynomial regression, neural network models, physics-informed hybrid residual modeling, and LASSO regression with polynomial feature expansion. The novelty of this study lies in evaluating permeability and breakthrough pressure within a common analytical framework while combining missing-data treatment, multicollinearity assessment, machine learning, physics-informed modeling, and interpretable empirical equation development. The outcomes are intended to identify the dominant controls on caprock sealing behavior and support safe CO₂ injection pressure design, preliminary site screening, and long-term storage integrity assessment in geological CO₂ sequestration systems.

2. Materials and Methods

2.1. Data Collection and Dataset Preparation

This study was conducted using a data-driven meta-analysis approach based on quantitative data compiled from published literature on caprock integrity in geological CO₂

sequestration. The collected sources included laboratory experiments, field-based investigations, numerical studies, and technical reports that presented caprock properties under defined geological, experimental, or reservoir conditions. Permeability-related data and supporting information, including permeability, porosity, lithological characteristics, depth, and caprock flow behavior, were extracted from studies focusing on caprock hydraulic properties, leakage potential, and multiscale caprock integrity assessment [2,10–14], with additional background support from caprock review studies [3]. Breakthrough-pressure-related data, including breakthrough pressure, confining pressure, temperature, depth, and porosity, were obtained mainly from experimental studies addressing capillary entry pressure, CO₂ breakthrough behavior, and sealing efficiency [7,15–17]. Additional studies on thermal, mechanical, numerical, wettability, risk-assessment, and field-scale caprock responses were used to support interpretation of the controlling mechanisms and storage behavior [5,6,18–25].

Studies were included if they reported at least one of the selected sealing-performance indicators (i.e., permeability or breakthrough pressure), together with one or more governing variables. The primary input variables considered were porosity (%), depth (m), temperature (°C), and confining pressure (MPa). Permeability (mD) and breakthrough pressure (MPa) were selected as output variables because they represent two complementary aspects of caprock sealing performance.

Because the compiled data were obtained from multiple independent studies, unit systems, reporting formats, and experimental conditions were first standardized. Permeability values were converted to millidarcies (mD), pressure values to megapascals (MPa), depth values to meters (m), and temperature values to degrees Celsius (°C). Duplicate entries, incomplete records with no usable output variable, and data points without clearly defined experimental or field conditions were excluded. The final compiled dataset covered a broad range of reservoir and caprock conditions, as summarized in Table 1.

Table 1. Summary of compiled dataset parameters, availability, and observed ranges.

Parameter	Unit	No. of Data Points	Observed Range	References
Permeability	mD	90	10 ^{−6} –10 ^{−1}	[2,5,10,11,13,14]
Porosity	%	70	0.5–25	[2,5,10,13,14]
Depth	m	65	500–4000	[2,4,10,11,13,19,22]
Breakthrough pressure	MPa	62	2–30	[11,15,17–19]
Confining pressure	MPa	55	5–60	[11,15,17–19]
Temperature	°C	55	15–90	[4,13,19,22]

2.2. Treatment of Missing Data

A major challenge in the compiled dataset was the presence of missing values, as most published studies did not report all selected variables simultaneously. Removing incomplete records would have reduced the dataset size and may have introduced selection bias. Therefore, MICE was used to estimate missing values while preserving statistical relationships among variables.

Prior to imputation, the missing data ratios were as follows: confining pressure (7.8%), temperature (38.9%), depth (27.8%), and porosity (22.2%). The application of MICE relies on the assumption that the data are Missing at Random (MAR), meaning the probability of a value being missing is related to the observed variables rather than the unobserved data itself. While MICE effectively preserves inter-variable dependencies, it

inherently introduces a degree of imputation uncertainty which propagates into the final predictive models.

MICE is an iterative imputation technique in which each incomplete variable is modeled conditionally using the remaining available variables in the dataset. The imputation procedure was performed over 100 iterations to improve convergence and stability of the estimated values. Following imputation, the number of complete usable records increased from 42 to 62 for breakthrough pressure modeling and from 58 to 90 for permeability modeling, representing improvements of 47.6% and 55.2%, respectively. This allowed a larger proportion of the available literature data to be retained while avoiding the information loss associated with complete-case deletion.

The quality of the MICE imputation was verified by comparing the density distributions of observed and imputed values for key variables, as shown in Figure 1, confirming that the imputed data maintains the statistical characteristics of the observed dataset.

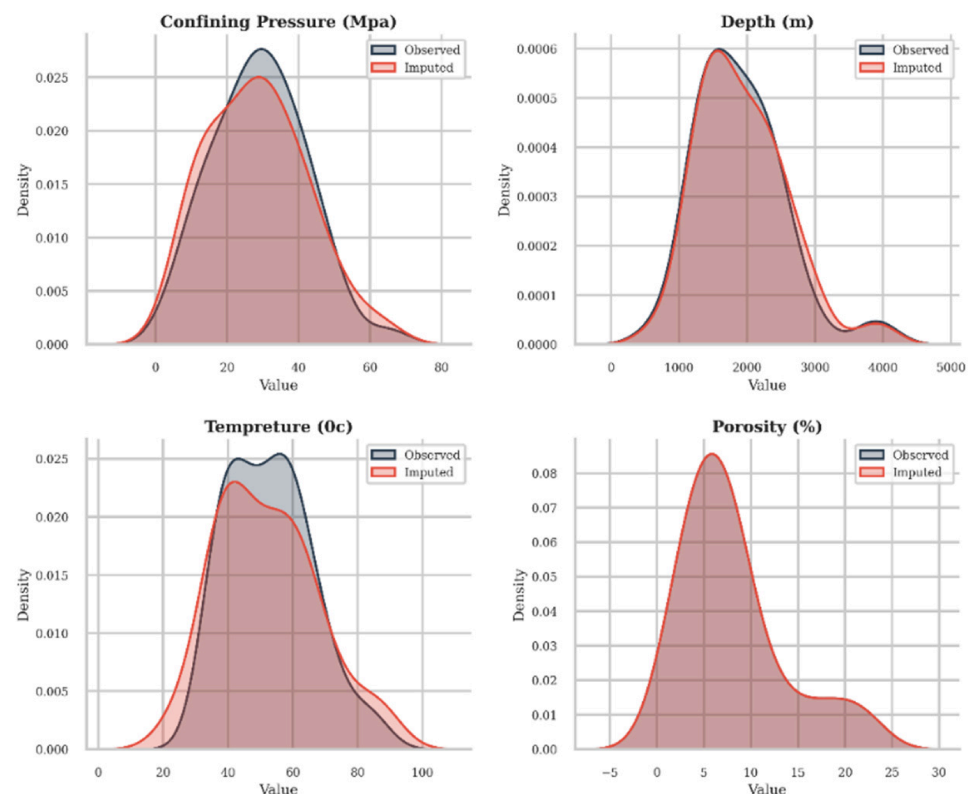


Figure 1. Kernel Density Estimation (KDE) comparison between original observed data and imputed data using MICE, confirming the statistical consistency of the imputation.

2.3. Permeability Analysis

Permeability analysis was carried out to evaluate the factors controlling fluid-flow potential through caprock formations. Because permeability values spanned several orders of magnitude, a logarithmic transformation was applied before statistical analysis. This transformation reduced the effect of extreme values and improved the suitability of the data for regression-based modeling. Both Pearson and Spearman correlation analyses were performed to examine relationships between log-transformed permeability and the selected input variables. Pearson correlation was used to identify linear associations, while Spearman correlation was used to evaluate monotonic relationships that may not be strictly linear. Spearman's correlation was prioritized for interpretation because caprock permeability is commonly influenced by nonlinear effects related to pore structure, stress history, lithological heterogeneity, and fracture development. Therefore, the permeability

analysis focused on identifying both direct and nonlinear trends between permeability and porosity, depth, and temperature.

2.4. Breakthrough Pressure Analysis

Breakthrough pressure analysis was conducted to evaluate the caprock sealing capacity against CO₂ entry. Breakthrough pressure was treated as a direct indicator of capillary sealing resistance because it represents the minimum pressure required for CO₂ to penetrate the water-saturated pore network of the caprock. The analysis considered confining pressure, temperature, depth, and porosity as the primary influencing variables.

Pearson and Spearman correlation analyses were used to quantify the strength and direction of relationships between breakthrough pressure and the selected input variables. Attention was given to confining pressure and temperature because these variables are closely related to reservoir stress conditions, pore throat closure, interfacial properties, and pressure-dependent sealing behavior. A comparative analysis was then performed to evaluate the relative influence of each variable on breakthrough pressure and to determine whether breakthrough pressure provides a more systematic sealing indicator than permeability.

2.5. Predictive Modeling Framework

Following the statistical analyses, predictive models were developed for both permeability and breakthrough pressure using a multi-model framework. Four modeling approaches were employed and compared: (1) Linear Regression as a baseline model, (2) Nonlinear regression models to capture nonlinear relationships and parameter interactions, (3) a basic Neural Network (NN) to explore deep pattern recognition within the dataset, and (4) a Physics-Informed Residual Model that integrates domain knowledge of caprock behavior with data-driven residual corrections to improve physical consistency of predictions. The multi-model framework was adopted to enable a structured comparison between conventional statistical approaches and more advanced hybrid techniques, ensuring that the relative advantages and limitations of each modeling strategy could be assessed within a unified analytical context. Model performance was evaluated using the coefficient of determination (R^2) and mean absolute error (MAE) to assess predictive accuracy and compare modeling approaches across both analyses. The methodological framework is illustrated in Figure 2.

To prevent severe information leakage, data preprocessing, including logarithmic transformations and feature scaling, was independently fitted on the training set and subsequently applied to the test set. However, due to the constrained sample size of the meta-analysis, MICE imputation was applied to the entire dataset prior to splitting to ensure convergence of the chained equations. This represents a standard and necessary methodological compromise in small-sample geo-mechanical studies to maintain statistical stability during missing-data estimation.

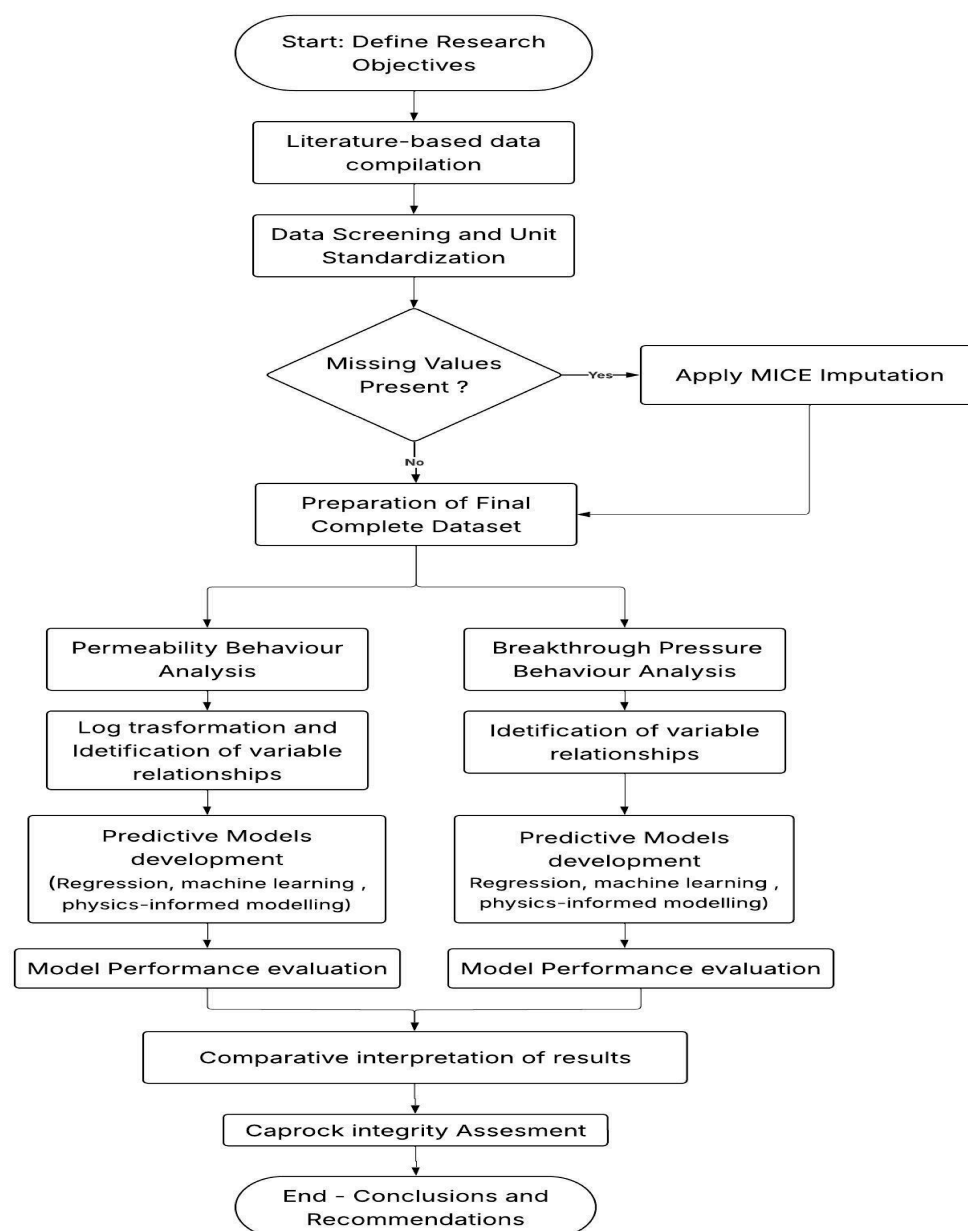


Figure 2. Methodological workflow used for the integrated data-driven assessment of caprock permeability and breakthrough pressure.

3. Results and Discussion

3.1. Permeability Analysis

Permeability analysis was conducted to evaluate fluid flow behavior within caprock formations using a dataset compiled from multiple literature sources. Due to the wide variation in permeability values spanning several orders of magnitude, a logarithmic transformation was applied prior to analysis to normalize the distribution. To examine variable relationships, both Pearson and Spearman correlation analyses were performed. Spearman correlation was prioritized given the non-linear nature of the data and is presented in Figure 3.

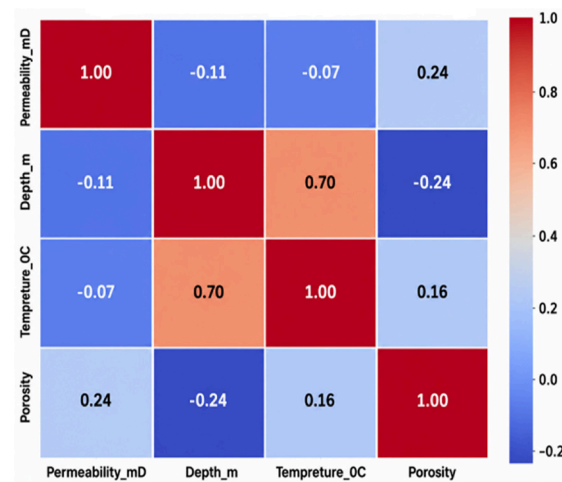


Figure 3. Spearman correlation matrix for permeability and influencing parameters. Red indicates a strong positive correlation (+1.0), while blue indicates a negative correlation, with color intensity reflecting the strength of the rank relationship.

The results indicate weak to moderate relationships overall. Porosity shows a moderate positive association with log-transformed permeability ($\rho \approx 0.24$), where ρ represents the Spearman correlation coefficient, which measures the strength and direction between variables. Depth and temperature exhibit weak correlations ($\rho \approx -0.11$ and -0.07 , respectively), suggesting limited direct linear influence within the dataset. These relationships are further illustrated through pairwise scatter plots in Figure 4.

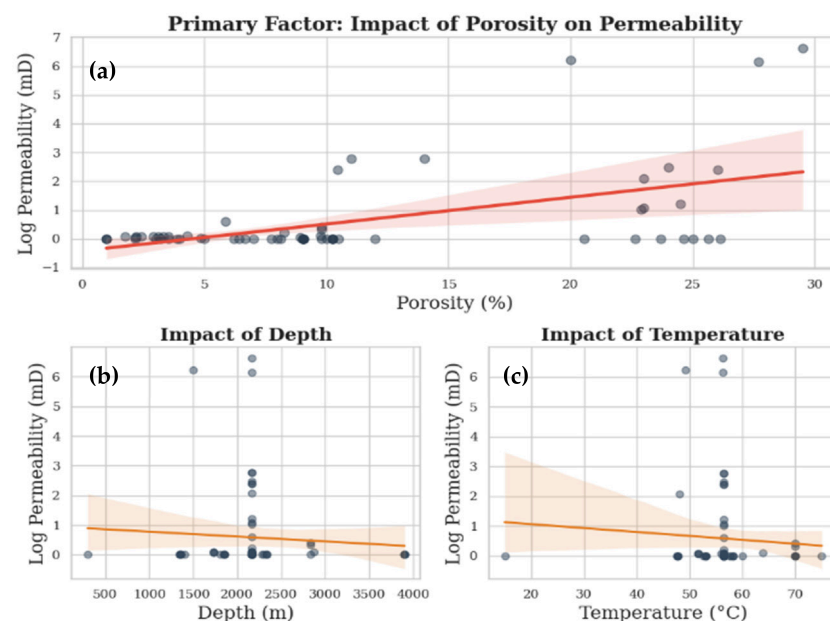


Figure 4. Pairwise scatter plots showing the relationships between log-transformed permeability and selected input parameters: (a) porosity, (b) depth, and (c) temperature. Fitted trend lines with confidence intervals are included to illustrate the general direction of each relationship.

The pairwise scatter plots show considerable dispersion, with no strong linear relationship between log-transformed permeability and the selected input parameters. This reflects the heterogeneous nature of caprock formations, where permeability is influenced not only by porosity, depth, and temperature, but also by lithology, pore connectivity, fracture networks, and stress history. Ordinary Least Squares (OLS) regression was first applied using porosity, depth, and temperature as predictors of log-transformed permeability. The model produced limited explanatory power, with $R^2 \approx 0.29$, confirming that permeability

could not be reliably predicted using a simple linear relationship. Among the selected predictors, porosity showed the most meaningful statistical contribution, supporting its role in controlling pore volume and flow capacity.

Depth and temperature showed strong multicollinearity, as indicated by high Variance Inflation Factor (VIF) values. This suggests that their individual regression coefficients should be interpreted cautiously because both variables are related to burial and thermal conditions. Polynomial regression did not improve model performance, indicating that the observed permeability variation is controlled by complex geological heterogeneity rather than by simple nonlinear trends alone. This weak predictive performance is consistent with previous caprock studies showing that permeability is highly site-specific and commonly governed by pore connectivity, lithological variation, fracture networks, and stress-dependent flow pathways [10,11]. SHapley Additive exPlanations (SHAP)-based sensitivity analysis using the Gradient Boosting model further indicated that depth and porosity were influential variables in permeability prediction. However, these effects should be interpreted as model-based sensitivity trends rather than universal physical relationships because permeability anomalies are highly site-specific and dependent on local geological conditions.

The Physics-Informed Residual Model provided the strongest predictive performance among all approaches applied to the permeability dataset. This model operates in two stages: a base physical model constructed using porosity as the primary input variable, grounded in the established relationship between pore volume and fluid flow capacity in fine-grained caprocks, followed by a data-driven residual correction that models the systematic deviations between base model predictions and observed permeability values as a function of depth and temperature. As shown in Figure 5, the base physical model achieved $R^2 = 0.28$, capturing the primary porosity-driven trend but leaving substantial variance unexplained due to the heterogeneous nature of the compiled dataset. The addition of the depth-temperature residual correction improved overall model performance to $R^2 = 0.77$, representing a 48.7% gain and confirming that the unexplained variance was systematic and physically meaningful rather than random noise. A neural network model achieved an intermediate R^2 of 0.424, with limited performance attributable to the small dataset size, which constrains the generalization capacity of purely data-driven approaches.

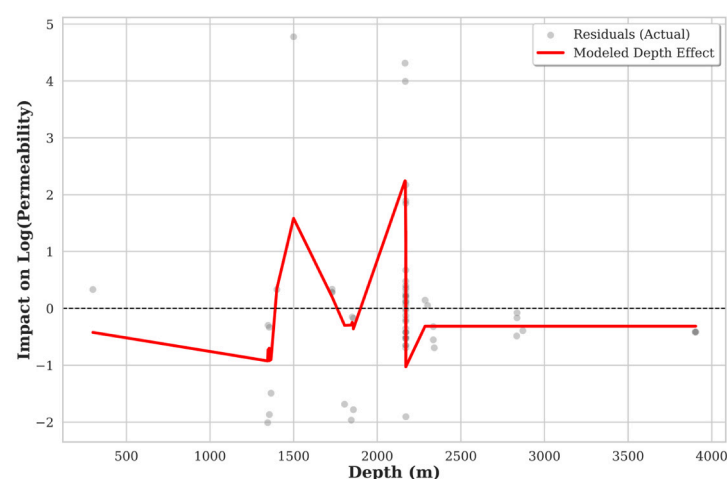


Figure 5. Depth-dependent residual response of log-transformed permeability after accounting for the primary porosity effect. Gray points represent the actual residuals, the red line indicates the modeled depth effect, and the dashed horizontal line denotes zero residual influence.

Residual analysis further identified two depth intervals, approximately 1400–1600 m and 2100–2300 m, that exhibited elevated permeability responses within the compiled

dataset. To describe these localized deviations, a dataset-specific depth-zone residual correction model was introduced by combining the background porosity-controlled trend with correction factors for the two identified depth intervals:

$$\text{Log } k = 0.0647 - 0.0021 \phi + 0.7670 I_A + 3.0659 I_B \quad (1)$$

where (Log k) is the log-transformed permeability response, ϕ is porosity (%), I_A is an indicator variable for the depth interval 2100–2300 m,

$$I_A = \begin{cases} 1, & 2100 \leq D \leq 2300 \text{ m} \\ 0, & \text{otherwise} \end{cases}$$

and I_B is an indicator variable for the depth interval 1400–1600 m,

$$I_B = \begin{cases} 1, & 1400 \leq D \leq 1600 \text{ m} \\ 0, & \text{otherwise} \end{cases}$$

where D denotes depth.

The correction factors indicate that the elevated permeability responses are not fully explained by porosity alone and may reflect localized fracture networks, lithological transitions, or site-specific stress and diagenetic histories represented in the source studies. Similar localized permeability deviations have been reported in caprock and fault-sealing studies, where fracture development, lithological transitions, and fault-related flow pathways can produce permeability responses that deviate from simple burial-compaction trends [11,13]. Because this depth-zone residual correction is heavily over-fitted to the anomalies of the compiled dataset, it cannot be presented as a generally applicable predictive model. Its robustness must be subjected to rigorous cross-validation against independent, external datasets before broader application. Overall, the results confirm that permeability does not follow a simple burial-compaction trend across heterogeneous caprock systems. Therefore, although porosity remains an important controlling parameter, permeability should be assessed together with complementary sealing indicators such as breakthrough pressure.

3.2. Breakthrough Pressure Analysis

Breakthrough pressure analysis was conducted as a direct measure of caprock sealing capacity. Compared to the permeability dataset, breakthrough pressure values were fully observed across all records, with only minor missing values, all of which were addressed using MICE prior to analysis. To examine relationships between variables, both Pearson and Spearman correlation analyses were performed, with Spearman's correlation presented in Figure 6 due to the non-linear characteristics of the dataset.

The results indicate stronger relationships compared to the permeability analysis. Temperature shows a strong positive correlation with breakthrough pressure ($\rho = 0.97$), followed by confining pressure ($\rho = 0.85$) and depth ($\rho = 0.78$), highlighting the dominant role of thermomechanical subsurface conditions in controlling sealing capacity. In contrast, porosity exhibits a moderate negative relationship ($\rho = -0.31$), suggesting that increased pore space reduces resistance to CO₂ penetration, consistent with capillary sealing theory. This behavior agrees with capillary-sealing studies showing that higher confining pressure can reduce pore-throat aperture and increase CO₂ entry pressure, whereas greater porosity can reduce capillary resistance [15,17]. These trends are further illustrated in the pairwise plots shown in Figure 7.

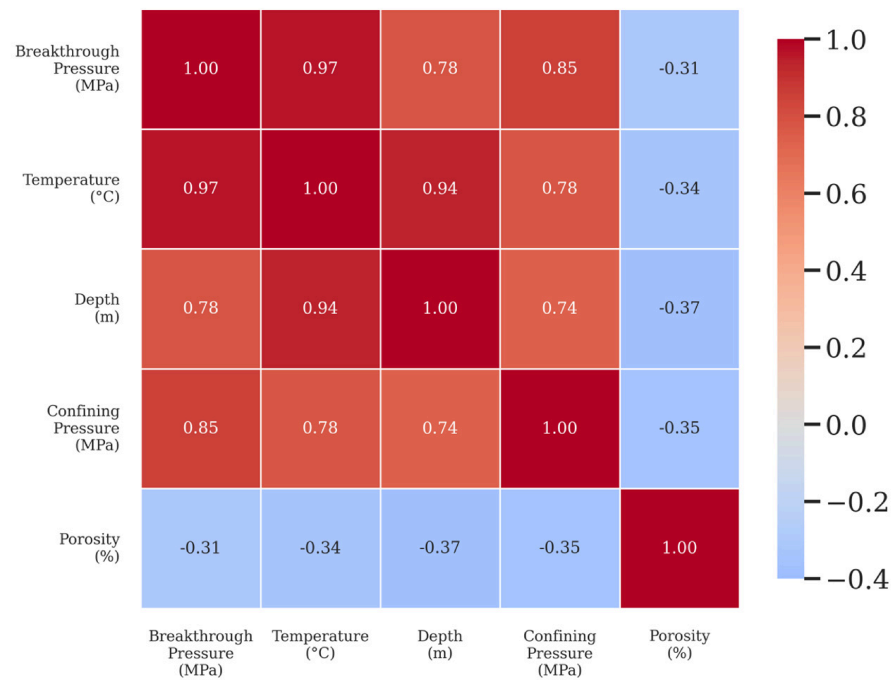


Figure 6. Spearman correlation matrix for breakthrough pressure and influencing parameters. Red indicates a strong positive correlation (+1.0), while blue indicates a negative correlation, with color intensity reflecting the strength of the rank relationship.

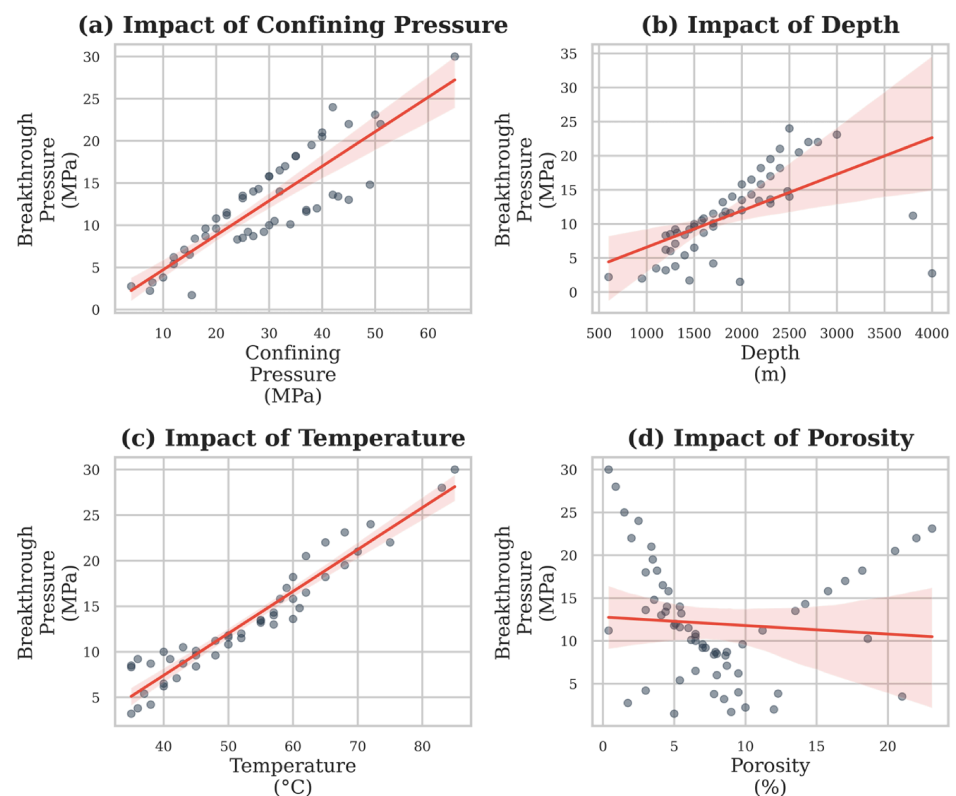


Figure 7. Pairwise scatter plots showing the relationships between breakthrough pressure and selected input parameters: (a) confining pressure, (b) depth, (c) temperature, and (d) porosity. Fitted trend lines with confidence intervals are included to illustrate the general direction of each relationship.

The scatter plots in Figure 7 confirm clear and consistent positive trends between breakthrough pressure and confining pressure, temperature, and depth, with significantly less dispersion compared to the permeability analysis. This indicates that breakthrough

pressure is more systematically governed by the selected parameters. The negative relationship with porosity is also visible, though with greater scatter reflecting natural geological heterogeneity.

Regression analysis further supports these observations. OLS regression using all predictors achieved high explanatory power ($R^2 = 0.975$), indicating that the selected variables capture most of the variability in breakthrough pressure. Confining pressure and temperature were the dominant predictors, while depth had an indirect influence due to its strong correlation with subsurface stress conditions ($VIF > 100$). Polynomial regression using confining pressure achieved $R^2 = 0.818$; however, higher-order terms were not statistically significant, confirming that the relationship between confining pressure and breakthrough pressure is predominantly linear within the observed range. SHAP-based feature importance analysis using the trained Gradient Boosting model was conducted to interpret the relative contribution of each input variable to breakthrough pressure prediction, as shown in Figure 8. Temperature produced the highest mean absolute SHAP value, followed by confining pressure, indicating that these two variables had the strongest influence on the model prediction. Porosity and depth showed comparatively lower mean absolute SHAP values, suggesting smaller individual contributions within the trained model. The dominant role of temperature is physically reasonable because reservoir temperature can influence CO_2 density, brine properties, interfacial behavior, and thermo-mechanical caprock response [5,22]. From a petrophysical perspective, elevated reservoir temperatures directly reduce CO_2 density and alter brine viscosity, which subsequently modifies the interfacial tension between the CO_2 and formation fluids, critically governing capillary entry resistance.

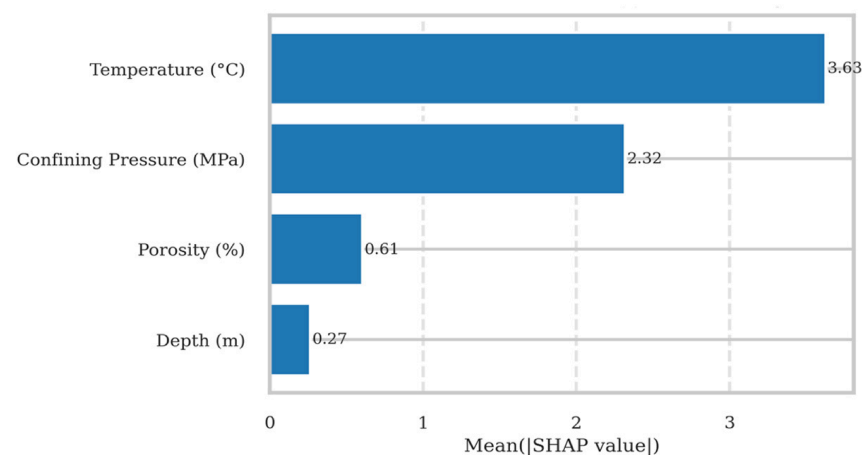


Figure 8. SHAP-based feature importance analysis for breakthrough pressure prediction using the trained Gradient Boosting model. The mean absolute SHAP value represents the average contribution of each input variable to the predicted breakthrough pressure.

The severe multicollinearity between depth and temperature ($VIF > 100$) complicates strict feature importance interpretation. In tree-based models, correlated features can share importance, meaning the high SHAP value for temperature likely absorbs some of the physical variance inherently driven by burial depth. Therefore, temperature and depth should be interpreted as a coupled thermo-mechanical driver rather than entirely independent variables.

The Physics-Informed Hybrid Model produced the strongest predictive performance, as illustrated in Figure 9. This model operates in two stages: a base physical model driven by depth and temperature as macro-scale geological drivers, followed by a data-driven residual correction incorporating confining pressure and porosity as pore-scale mechanical parameters. The base physical model achieved moderate performance of

$R^2 = 0.688$, capturing the broad depth and temperature-driven trends in breakthrough pressure but leaving residual variance attributable to stress and pore-structure effects. The addition of the residual correction significantly improved predictive accuracy to $R^2 = 0.997$, demonstrating that combining macro-scale geological drivers with pore-scale mechanical parameters effectively captures the underlying behavior of breakthrough pressure within the compiled dataset. It is critical to note that the R^2 of 0.997 achieved by the hybrid model represents training performance only. Given the limited dataset size, this near-perfect fit may indicate overfitting. Therefore, this hybrid model serves primarily to demonstrate the theoretical alignment of physical and statistical parameters, rather than acting as a highly generalizable predictive tool.

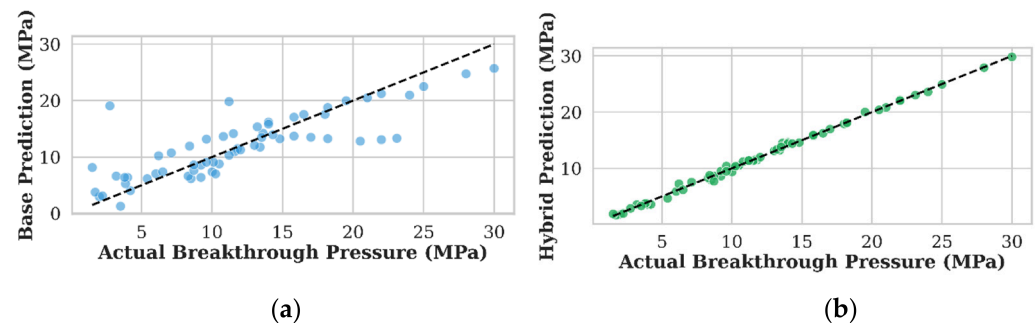


Figure 9. Comparison between actual and predicted breakthrough pressure values for (a) the base physical model and (b) the physics-informed hybrid model. The dashed diagonal line represents the ideal 1:1 prediction line, where predicted values are equal to actual values.

Given the limited dataset size and the poor generalization capacity observed with the neural network approach, LASSO-based polynomial regression was adopted as a more stable and interpretable alternative for deriving an empirical predictive equation for breakthrough pressure. The LASSO model incorporated all four input variables: confining pressure (C), temperature (T), depth (D), and porosity (φ), and selected the most influential polynomial terms through regularization, producing the following empirical equation:

$$P_b = 2.1445 + C(0.0026 C + 0.0043 T - 0.1714) + T(0.5541 + 0.0121 \varphi) + \varphi(0.0941 \varphi + 0.00001 D - 2.4398) - 0.000005 D^2 \quad (2)$$

where P_b is the breakthrough pressure (MPa); C is the confining pressure (MPa); T is the temperature ($^{\circ}\text{C}$); D is depth (m); and φ is porosity (%). The model achieved a training R^2 of 0.78, a test R^2 of 0.71, and a MAE of 2.3087 MPa, demonstrating acceptable predictive accuracy and reasonable generalizability relative to the training performance. The moderate gap between training and test R^2 confirms that the model generalizes adequately without significant overfitting, making it more suitable for the limited dataset size than the neural network approach. To ensure the safe application of this empirical equation for preliminary site screening, it must strictly be applied within the parameter ranges of the training dataset: Porosity (0.5–25%), Depth (500–4000 m), Temperature (15–90 $^{\circ}\text{C}$), and Confining Pressure (5–60 MPa). Extrapolating beyond these bounds removes the statistical validity of the correlation. This equation provides a practical and interpretable tool for preliminary breakthrough pressure estimation based on measurable subsurface parameters. Furthermore, a leverage analysis using Cook's Distance was conducted to ensure model stability (Figure 10), confirming that the empirical coefficients were not disproportionately driven by isolated high-leverage observations in the compiled dataset.

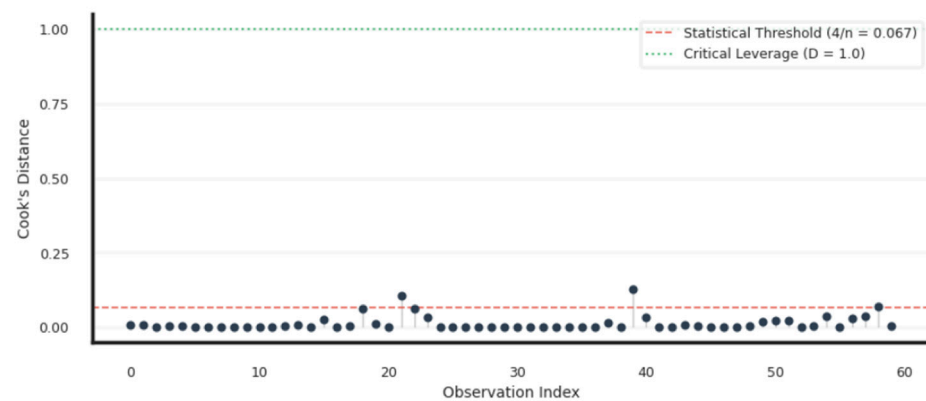


Figure 10. Cook's distance leverage analysis. Stem plot evaluating model stability for the training dataset.

Overall, the breakthrough pressure analysis demonstrates that caprock sealing capacity is strongly and consistently governed by confining pressure and temperature, with depth playing a secondary but meaningful role. Compared to permeability, breakthrough pressure exhibits more predictable and systematic relationships with the selected variables, confirming its reliability as a primary indicator of caprock sealing performance in geological CO₂ sequestration systems.

3.3. Comparative Analysis

The comparative analysis shows a clear difference between the predictability of permeability and breakthrough pressure. Permeability exhibited weak performance in OLS and polynomial regression, with R^2 values of approximately 0.29, reflecting the influence of lithological heterogeneity, fracture networks, and local pore-structure variability. In contrast, breakthrough pressure showed substantially stronger regression performance, with OLS achieving $R^2 = 0.975$, indicating that the selected variables captured the major controls on capillary sealing resistance.

The neural network models did not outperform the simpler approaches for either parameter, suggesting that the available dataset was insufficient for purely data-driven deep learning models. This limitation was more evident for breakthrough pressure, where the neural network produced lower performance than the physics-informed and regularized regression approaches.

The physics-informed models provided the strongest overall performance for both parameters. For permeability, the residual model improved R^2 from 0.281 to 0.768, indicating that depth- and temperature-dependent residual corrections captured systematic geological effects not explained by porosity alone. For breakthrough pressure, the physics-informed hybrid model improved R^2 from 0.688 to 0.997, demonstrating that combining macro-scale geological drivers with pore-scale mechanical parameters provides an effective representation of sealing behavior.

Overall, breakthrough pressure was more predictable than permeability because it directly represents capillary entry resistance, whereas permeability is more strongly affected by local heterogeneity and fracture-controlled flow paths. This interpretation is consistent with previous caprock integrity studies, which emphasize that permeability represents potential leakage through connected flow paths, while breakthrough pressure provides a more direct measure of capillary entry resistance and sealing efficiency [15,21].

A consolidated summary of the key quantitative outcomes across both analyses, including model performance and physical interpretation, is presented in Table 2.

Table 2. Summary of results.

Model	Permeability	Breakthrough Pressure	Interpretation
OLS Regression	$R^2 = 0.290$	$R^2 = 0.975$	Linear baseline
Polynomial Regression	$R^2 \approx 0.288\text{--}0.289$	$R^2 = 0.680\text{--}0.818$	Polynomial extensions did not resolve nonlinear permeability behavior
Neural Network	$R^2 = 0.414$	$R^2 = 0.231$	Limited by small dataset size; overfitting risk
LASSO Regression	-	Training $R^2 = 0.78$, Test $R^2 = 0.71$	Stable empirical equation derived for breakthrough pressure
Base Physical Model	$R^2 = 0.281$	$R^2 = 0.688$	Captures first-order physical trend only
Physics-Informed Hybrid Model	$R^2 = 0.768$	$R^2 = 0.997$	Best training data performance across both analyses

3.4. Uncertainty Analysis and Model Limitations

While the proposed data-driven framework offers strong predictive capabilities for preliminary site screening, it relies on macroscopic variables that necessitate physical simplifications. A primary limitation is the absence of microscopic pore structure parameters in the compiled literature, which restricts a complete mechanistic interpretation of capillary resistance and contributes to observed data dispersion. Furthermore, the dataset integrates various fine-grained caprocks without explicit mineralogical classification, introducing systematic predictive bias since varying clay and carbonate contents fundamentally alter rock wettability and interfacial tension. The models also omit sedimentary facies constraints, representing an additional source of unquantified uncertainty. Finally, substituting simple burial depth and temperature for complex diagenetic and tectonic evolution, which can induce matrix-bypassing fractures, means the applicability of these empirical equations must be strictly constrained to matrix-dominated, fracture-free systems. Ultimately, applying these models beyond the current parameter ranges carries unvalidated extrapolation risks, necessitating rigorous, localized characterizations of mineralogy and diagenetic history for future site-specific engineering applications.

4. Conclusions

This study evaluated the role of caprock integrity in geological CO₂ sequestration using a data-driven meta-analysis of permeability and breakthrough pressure as the two main indicators of sealing performance. The findings show that caprock behavior cannot be fully understood using a single parameter, as sealing capacity is controlled by a combination of geological, hydraulic, thermal, and mechanical factors.

Permeability analysis showed weak linear predictability, with the OLS model producing an R^2 of 0.29. This indicates that permeability is highly sensitive to geological heterogeneity, fracture condition, lithological variation, and pore connectivity, making it difficult to predict using simple linear relationships. Although porosity was identified as the most consistent controlling parameter for permeability, it alone was insufficient to

explain the full variation observed in the compiled dataset. The application of a Physics-Informed Residual Model improved permeability prediction to $R^2 = 0.77$, demonstrating that combining a porosity-based physical trend with depth-related residual correction can better represent complex permeability behavior.

In contrast, breakthrough pressure showed stronger and more systematic relationships with the selected variables. It was positively associated with temperature, confining pressure, and depth, whereas porosity showed a negative relationship consistent with capillary-sealing theory. The OLS model achieved high training performance ($R^2 = 0.975$), suggesting that the selected variables explain most of the observed variation. However, the presence of multicollinearity among depth, temperature, and confining pressure means that individual regression coefficients should be interpreted carefully. The Physics-Informed Hybrid Model achieved the highest training performance ($R^2 = 0.997$), confirming that integrating broad subsurface conditions with caprock-specific properties can significantly improve model fit. A LASSO-derived empirical equation subsequently achieved a validated test R^2 of 0.78 with a mean absolute error of 2.31 MPa on a held-out test set, providing the most generalizable predictive correlation for practical application.

The use of MICE imputation improved dataset usability by enabling retention of 90 permeability records and 62 breakthrough pressure records despite incomplete reporting across literature sources. Overall, breakthrough pressure is a more reliable primary indicator of caprock sealing capacity than permeability, because it directly represents the pressure required for CO₂ to penetrate the caprock seal. Permeability remains important as a complementary hydraulic leakage-risk indicator but should not be used alone for caprock integrity assessment.

From an engineering perspective, the predicted breakthrough pressure can be utilized to establish operational safety margins for CO₂ injection. Specifically, the maximum allowable injection pressure at the storage site must be strictly maintained below the predicted capillary entry pressure, typically incorporating a conservative safety factor to account for geological uncertainties and localized matrix heterogeneity. Similarly, the predicted permeability serves as a threshold indicator for site screening; caprock systems exceeding permeability targets should trigger intensified seismic and pressure monitoring protocols to ensure long-term containment integrity.

From a practical perspective, preliminary CO₂ storage site screening should prioritize the measurement of breakthrough pressure, confining pressure, porosity, caprock continuity, lithology, and fracture conditions. The study also demonstrates that physics-informed modeling provides a useful framework for combining geological understanding with data-driven correction; however, independent validation is required before applying the models in operational storage projects.

While the random partitioning of the compiled dataset successfully establishes internal model consistency, it does not fully substitute for independent external validation. We acknowledge the inherent risk of latent biases within aggregated literature sources as a recognized constraint of meta-analytical frameworks. Consequently, the proposed predictive models are most appropriately deployed as foundational screening tools, intended to complement rather than replace rigorous, site-specific petrophysical investigations. To systematically isolate and quantify the impact of source-specific biases, future research should prioritize a 'leave-one-source-out' cross-validation methodology, alongside validation against field data from active injection operations.

Future research should focus on developing lithology-specific datasets by separating shale, mudstone, claystone, evaporites, and carbonate seals, allowing the influence of each rock type on permeability, breakthrough pressure, and mechanical strength to be assessed more accurately. Additional parameters such as clay mineral content, fracture density,

bedding orientation, pore-throat size, and injection pressure should also be incorporated. Finally, the proposed modeling framework should be validated using independent laboratory datasets and field data from active CO₂ storage projects to improve its reliability for long-term containment assessment.

Author Contributions: Conceptualization and methodology: R.G.C.T. and K.H.S.M.S.; software, validation, formal analysis, and investigation: R.G.C.T.; writing—original draft preparation, visualization: R.G.C.T.; review and editing, supervision, project administration, and funding acquisition: K.H.S.M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the University of Moratuwa Senate Research Committee SRC grant (SRC/LT/2025/11).

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

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Abbreviations

The following abbreviations and symbols are used in this manuscript:

CCS	Carbon Capture and Storage
CO ₂	Carbon Dioxide
MICE	Multiple Imputation by Chained Equations
OLS	Ordinary Least Squares
VIF	Variance Inflation Factor
SHAP	SHapley Additive exPlanations
LASSO	Least Absolute Shrinkage and Selection Operator
NN	Neural Network
R ²	Coefficient of Determination
MAE	Mean Absolute Error

References

1. Cao, C.; Zhu, H.; Hou, Z. Advances in Carbon Capture, Utilization and Storage (CCUS). *Energies* **2024**, *17*, 4784. [\[CrossRef\]](#)
2. Sazali, W.M.L.b.; Busch, A. Determination of Caprock Porosity and Permeability Using Drill Cuttings and Core Samples of a Potential CO₂ Storage Site—A Case Study. In Proceedings of the Abu Dhabi International Petroleum Exhibition & Conference (ADIPEC), Abu Dhabi, United Arab Emirates, 31 October–3 November 2022. [\[CrossRef\]](#)
3. Shukla, R.; Ranjith, P.; Haque, A.; Choi, S.-K.X. A review of studies on CO₂ sequestration and caprock integrity. *Fuel* **2010**, *89*, 2651–2664. [\[CrossRef\]](#)
4. Wu, H.; Dai, Q.; Wang, R.; Zhou, Y.; Zhang, Y. Evolution of Caprock Sealing Capacity Under CO₂–Mechanical Coupling in Geological Carbon Storage. *Processes* **2025**, *13*, 3863. [\[CrossRef\]](#)
5. Iglaier, S.; Pentland, C.H.; Busch, A. CO₂ wettability of seal and reservoir rocks and the implications for carbon geo-sequestration. *Water Resour. Res.* **2015**, *51*, 729–774. [\[CrossRef\]](#)
6. Bashir, A.; Ali, M.; Patil, S.; Aljawad, M.S.; Mahmoud, M.; Al-Shehri, D.; Hoteit, H.; Kamal, M.S. Comprehensive review of CO₂ geological storage: Exploring principles, mechanisms, and prospects. *Earth-Sci. Rev.* **2024**, *249*, 104672. [\[CrossRef\]](#)
7. Espinoza, D.N.; Awad, M.M.; Kim, K.; Yu, S. Experimental Studies on Caprock Integrity for Safe Geological Carbon Storage. In Proceedings of the 17th Greenhouse Gas Control Technologies Conference (GHGT-17), Calgary, AB, Canada, 20–24 October 2024. [\[CrossRef\]](#)
8. Dooley, J.J. Estimating the Supply and Demand for Deep Geologic CO₂ Storage Capacity over the Course of the 21st Century: A Meta-analysis of the Literature. *Energy Procedia* **2013**, *37*, 5141–5150. [\[CrossRef\]](#)
9. Sathsarani, H.B.S.; Sampath, K.H.S.M.; Ranathunga, A.S. Utilization of fly ash-based geopolymer for well cement during CO₂ sequestration: A comprehensive review and a meta-analysis. *Gas Sci. Eng.* **2023**, *113*, 204974. [\[CrossRef\]](#)

10. Rizzo, R.E.; Inskip, N.F.; Fazeli, H.; Betlem, P.; Bisdom, K.; Kampman, N.; Snippe, J.; Senger, K.; Doster, F.; Busch, A. Modelling geological CO₂ leakage: Integrating fracture permeability and fault zone outcrop analysis. *Int. J. Greenh. Gas Control* **2024**, *133*, 104105. [\[CrossRef\]](#)
11. Zappone, A.; Rinaldi, A.P.; Grab, M.; Wenning, Q.C.; Roques, C.; Madonna, C.; Obermann, A.C.; Bernasconi, S.M.; Brennwald, M.S.; Kipfer, R.; et al. Fault sealing and caprock integrity for CO₂ storage: An in situ injection experiment. *Solid Earth* **2021**, *12*, 319–343. [\[CrossRef\]](#)
12. Mi, L.; Guo, Y.; Li, Y.; Zeng, D.; Lu, C.; Zhang, G. Evaluation of the dynamic sealing performance of cap rocks of underground gas storage under multi-cycle alternating loads. *Energy Geosci.* **2024**, *5*, 100319. [\[CrossRef\]](#)
13. Trujillo, N.; Rose-Coss, D.; Heath, J.E.; Dewers, T.A.; Ampomah, W.; Mozley, P.S.; Cather, M. Multiscale Assessment of Caprock Integrity for Geologic Carbon Storage in the Pennsylvanian Farnsworth Unit, Texas, USA. *Energies* **2021**, *14*, 5824. [\[CrossRef\]](#)
14. Jia, S.; Wen, C.; Fu, X.; Liu, T.; Xi, Z. A Caprock Evaluation Methodology for Underground Gas Storage in a Deep Depleted Gas Reservoir: A Case Study for the X9 Lithologic Trap of Langgu Sag, Bohai Bay Basin, China. *Energies* **2022**, *15*, 4351. [\[CrossRef\]](#)
15. Espinoza, D.N.; Santamarina, J.C. CO₂ breakthrough—Caprock sealing efficiency and integrity for carbon geological storage. *Int. J. Greenh. Gas Control* **2017**, *66*, 218–229. [\[CrossRef\]](#)
16. Fentaw, J.W.; Emadi, H.; Hussain, A.; Fernandez, D.M.; Thiyagarajan, S.R. Geochemistry in Geological CO₂ Sequestration: A Comprehensive Review. *Energies* **2024**, *17*, 5000. [\[CrossRef\]](#)
17. Gao, S.; Wei, N.; Li, X.; Wang, Y.; Wang, Q. Cap Rock CO₂ Breakthrough Pressure Measurement Apparatus and Application in Shenhua CCS Project. *Energy Procedia* **2014**, *63*, 4766–4772. [\[CrossRef\]](#)
18. Ringrose, P.S.; Mathieson, A.S.; Wright, I.W.; Selama, F.; Hansen, O.; Bissell, R.; Saoula, N.; Midgley, J. The in Salah CO₂ Storage Project: Lessons Learned and Knowledge Transfer. *Energy Procedia* **2013**, *37*, 6226–6236. [\[CrossRef\]](#)
19. Arts, R.; Eiken, O.; Chadwick, A.; Zweigel, P.; van der Meer, L.; Zinszner, B. Monitoring of CO₂ injected at Sleipner using time-lapse seismic data. *Energy* **2004**, *29*, 1383–1392. [\[CrossRef\]](#)
20. Shi, J.-Q.; Korre, A.; Chen, D.; Govindan, R.; Durucan, S. A methodology for CO₂ storage system risk and uncertainty assessment. *Energy Procedia* **2014**, *63*, 4750–4757. [\[CrossRef\]](#)
21. Chen, B.; Li, Q.; Tan, Y. Caprock sealing for geologic CO₂ storage: Research advances, challenges and prospects. *J. Rock Mech. Geotech. Eng.* **2025**, *18*, 335–363. [\[CrossRef\]](#)
22. Gor, G.Y.; Prévost, J.H. Effect of CO₂ Injection Temperature on Caprock Stability. *Energy Procedia* **2013**, *37*, 3727–3732. [\[CrossRef\]](#)
23. Hawez, H.K.; Bakir, H.; Jamal, K.; Kakakhan, M.; Hussein, K.; Omar, M. Numerical Investigation of CO₂ Injection Effects on Shale Caprock Integrity: A Case Study of Opalinus Clay. *Gases* **2025**, *5*, 15. [\[CrossRef\]](#)
24. Ahmadinia, M.; Shariatipour, S.M. Analysing the role of caprock morphology on history matching of Sleipner CO₂ plume using an optimisation method. *Greenh. Gases Sci. Technol.* **2020**, *10*, 1077–1097. [\[CrossRef\]](#)
25. Kalam, S.; Olayiwola, T.; Al-Rubaii, M.M.; Amaechi, B.I.; Jamal, M.S.; Awotunde, A.A. Carbon dioxide sequestration in underground formations: Review of experimental, modeling, and field studies. *J. Pet. Explor. Prod. Technol.* **2020**, *11*, 303–325. [\[CrossRef\]](#)

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.