

Disentangling Operating Conditions from Intrinsic Adsorbent Properties in CO₂ Capture: An Explainable Machine Learning Study Across Four Adsorbent Categories

Rahma H. Thabet

`rahma.hussien@sh-eng.menofia.edu.eg`

Menoufia University

Research Article

Keywords: CO₂ adsorption, Carbon capture, Adsorbent materials, Explainable machine learning, Artificial intelligence, Data-driven modeling, Adsorption capacity

Posted Date: September 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10878910/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Disentangling Operating Conditions from Intrinsic Adsorbent Properties in CO₂ Capture: An Explainable Machine Learning Study Across Four Adsorbent Categories

Rahma H. Thabet^{1*}

¹*Basic Engineering Science Department, Faculty of Engineering, Menoufia University, Shebin El-Kom, Egypt*

*Corresponding author: Tel: +2 0482221549, Fax: +2 0482235695; E-mail: rahma.hussien@sh-eng.menoufia.edu.eg, ORCID ID: [0009-0004-1810-1763](https://orcid.org/0009-0004-1810-1763)

Abstract

CO₂ adsorption performance depends jointly on intrinsic material properties and experimental operating conditions, complicating adsorbent comparisons across the literature. Here, a curated database of 398 CO₂ adsorption measurements spanning four adsorbent categories, carbon-based, organic porous, inorganic porous, and metal oxide adsorbents, was analyzed using exploratory data analysis and explainable machine learning to identify the key descriptors controlling adsorption capacity. BET surface area, pore volume, temperature, CO₂ partial pressure, and adsorbent family were identified as the most informative predictors. An optimized Gradient Boosting model achieved the highest predictive accuracy ($R^2 = 0.930$, RMSE = 3.567 mmol g⁻¹, MAE = 1.496 mmol g⁻¹), while a Random Forest model was used for the subsequent SHAP-based explainability analysis. CO₂ partial pressure dominated predictions across the complete dataset, but its influence receded once pressure effects were minimized, revealing more balanced contributions from BET surface area, pore volume, temperature, and adsorbent family. The central contribution of this work is the quantitative demonstration that literature-derived adsorbent databases are confounded by an unequal distribution of testing pressures across material families, such that family-level performance rankings computed without accounting for this confound can misrepresent intrinsic material behaviour. More consistent reporting of micropore characteristics is recommended to improve future data-driven prediction of CO₂ adsorption.

Keywords: CO₂ adsorption; Carbon capture; Adsorbent materials; Explainable machine learning; Artificial intelligence; Data-driven modeling; Adsorption capacity

Introduction

The continued increase in atmospheric carbon dioxide (CO₂) concentration remains one of the most significant challenges to achieving global climate mitigation targets. Despite the rapid expansion of renewable energy technologies, fossil fuels continue to supply a substantial

proportion of global energy demand, resulting in persistent CO₂ emissions from power generation, cement production, steel manufacturing, and other energy-intensive industries. According to the International Energy Agency (IEA), carbon capture technologies will play an indispensable role in achieving net-zero emissions by enabling substantial reductions in industrial CO₂ emissions from sectors where complete electrification remains technically or economically challenging (Abdi et al.; Sovacool et al., 2024). Consequently, the development of efficient, scalable, and economically viable CO₂ capture technologies has become a global research priority. Among the available carbon capture technologies, adsorption-based separation has attracted considerable attention because of its relatively low energy requirement, operational flexibility, simple process configuration, and potential for cyclic regeneration compared with conventional chemical absorption processes (Zhu et al.). The performance of adsorption systems depends primarily on the physicochemical characteristics of the adsorbent, including surface area, pore structure, surface chemistry, and adsorption affinity, together with operating parameters such as temperature and pressure. Consequently, extensive efforts have been devoted to developing advanced porous adsorbents with improved adsorption capacity, selectivity, regeneration stability, and long-term durability (Ibrahim and Hussein). The development of advanced porous adsorbents has become a central research focus in adsorption-based carbon capture, resulting in a wide variety of materials with distinct structural characteristics, adsorption mechanisms, and physicochemical properties. Among these, carbon-based adsorbents (e.g., activated carbons and porous carbons) are widely recognized for their high specific surface area, excellent chemical stability, moisture tolerance, and low production cost (Sai Bhargava Reddy et al., 2021). Organic porous materials, including covalent organic frameworks (COFs) and porous organic polymers (POPs), have attracted increasing attention because of their highly tunable pore architecture and surface functionality (Geng et al.), enabling enhanced CO₂ adsorption through rational molecular design (Jia et al.). Inorganic porous adsorbents, such as zeolites and mesoporous silica, offer well-defined pore structures and excellent thermal stability, although their adsorption performance is often influenced by framework composition and moisture sensitivity (Kumar et al., 2018). Metal oxide adsorbents exhibit fundamentally different adsorption mechanisms, relying predominantly on chemisorption through reversible carbonate formation, making them particularly attractive for intermediate- and high-temperature carbon capture applications (Gunathilake and Jaroniec, 2016). Although these material families have been extensively investigated, identifying the most promising adsorbents remains challenging because adsorption capacities reported in the literature are generated under substantially different experimental conditions. Variations in adsorption pressure, temperature, gas composition, particle characteristics, activation procedures, and measurement protocols frequently exceed the intrinsic differences among adsorbent materials themselves (Zentou et al., 2025). Existing comparative studies commonly summarize adsorption capacities or rank materials according to their maximum reported performance, implicitly assuming that measurements obtained under different operating conditions are directly comparable. However, adsorption is governed by the combined influence of intrinsic structural descriptors, including surface area and pore characteristics, together with external operating variables such as pressure and temperature

(Gautam et al., 2023). The increasing availability of adsorption data has accelerated the application of machine learning (ML) in carbon capture research for predicting CO₂ adsorption performance, screening promising adsorbents, and guiding material design (Alatefi et al.). Compared with conventional empirical correlations, ML can capture complex nonlinear relationships among structural descriptors, operating conditions, and adsorption capacity without requiring explicit mechanistic assumptions. Consequently, numerous studies have reported accurate prediction of CO₂ adsorption using ensemble learning and deep learning algorithms (Deng et al.; Guo et al.). Despite these advances, most existing studies have primarily focused on improving predictive accuracy, while comparatively little attention has been given to interpreting how individual material descriptors and operating conditions influence model predictions. This limitation reduces the scientific value of ML models, particularly when heterogeneous literature datasets are used, where variations in experimental conditions may obscure intrinsic structure–property relationships (Al-Sakkari et al.). Explainable artificial intelligence (XAI) addresses this challenge by providing quantitative interpretation of feature contributions and enabling physically meaningful analysis of complex models. Techniques such as SHapley Additive exPlanations (SHAP) and Partial Dependence Plots (PDPs) have emerged as powerful tools for linking machine learning predictions with adsorption mechanisms, thereby improving model transparency and supporting more reliable scientific interpretation (Alatefi et al.; Wang et al.; Yuan et al.; Zhu et al.).

1.2. Research Gap and Objectives

Although machine learning has been increasingly applied to predict CO₂ adsorption performance, most studies have focused primarily on maximizing prediction accuracy using limited material classes or homogeneous experimental datasets. Comparatively little attention has been devoted to understanding how variations in operating conditions influence the interpretation of adsorption performance across structurally diverse adsorbent families. Furthermore, the lack of interpretable machine learning frameworks has limited the ability to distinguish the relative contributions of intrinsic material descriptors and experimental variables, reducing the scientific insights that can be extracted from literature-derived adsorption data.

To address these challenges, this study aims to develop an explainable machine learning framework for the analysis of CO₂ adsorption across four major adsorbent categories. The specific objectives are to:

1. **Construct and curate** a literature database of CO₂ adsorption measurements representing carbon-based adsorbents, organic porous materials, inorganic porous adsorbents, and metal oxide adsorbents.
2. **Identify the key physicochemical descriptors and operating variables** governing CO₂ adsorption capacity through exploratory data analysis and systematic feature selection.

3. **Develop and evaluate explainable machine learning models** for accurate prediction of CO₂ adsorption capacity and interpret the influence of individual descriptors using SHapley Additive exPlanations (SHAP) and Partial Dependence Plots (PDPs).
4. **Disentangle the effects of operating conditions and intrinsic material properties** to establish more reliable structure–property relationships and provide recommendations for improving future adsorption databases and data-driven adsorbent design.

2. Material and Methods

2.1. Database Construction

A literature-derived database was established by collecting experimental CO₂ adsorption data from peer-reviewed articles published between 2000 and 2025 using the Scopus, Web of Science, and ScienceDirect databases. Only experimental studies reporting CO₂ adsorption capacity together with the corresponding operating conditions and material characteristics were included. Following quality control, including duplicate removal and exclusion of incomplete observations, the final curated dataset comprised 398 adsorption measurements representing four major adsorbent categories: carbon-based adsorbents, organic porous materials, inorganic porous adsorbents, and metal oxide adsorbents. The overall workflow for database construction and data screening is illustrated in Table 1 and Fig. 1. To facilitate comparative analysis and machine learning, the individual adsorbent materials collected from the literature were grouped into four major adsorbent categories according to their chemical composition and structural characteristics (Table 2).

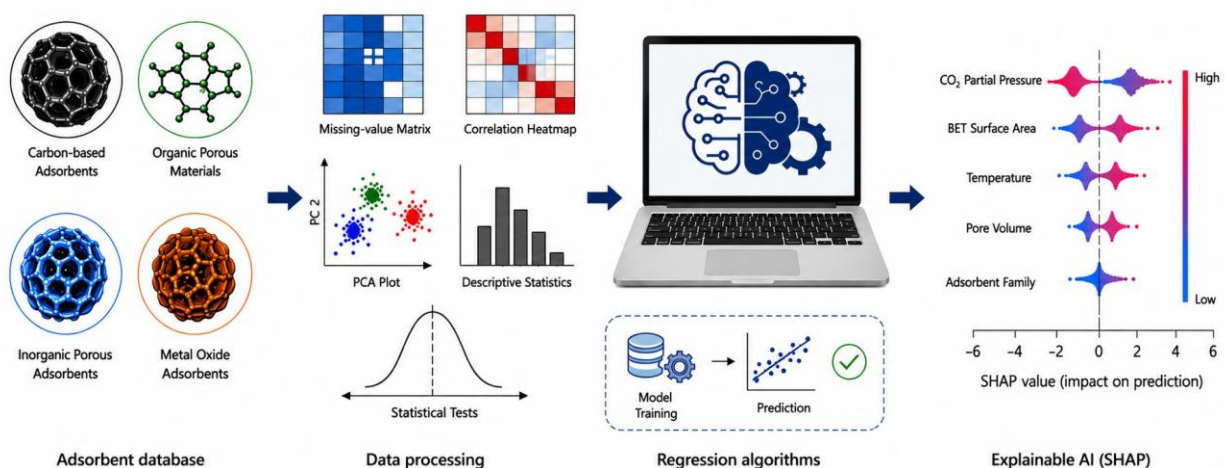


Figure 1. Workflow for literature screening, database construction, and machine learning analysis.

Table 1 Characteristics of the curated database.

Item	Value
Total observations	398
Main families	4
Material categories	11
Numerical features	6
Categorical features	2 (Category, Main Family)
Target variable	CO ₂ Adsorption Capacity (mmol g ⁻¹)

Table 2. Classification of the adsorbent materials into the four main families used in the present study.

Main Family	Constituent Categories
Carbon-based Adsorbents	Activated Carbon, Carbon Nanotubes, Porous Carbon
Organic Porous Materials	Covalent Organic Frameworks (COFs), Porous Organic Polymers (POPs)
Inorganic Porous Adsorbents	Zeolites, Mesoporous Silica
Metal Oxide Adsorbents	Magnesium Oxide (MgO), Calcium Oxide (CaO), Lithium Zirconate (Li ₂ ZrO ₃), Porous Oxides

This classification reduced the complexity of the database while preserving the physicochemical similarities among related adsorbent materials, enabling systematic comparisons across the major adsorbent families.

2.2. Data Preprocessing

Prior to model development, the curated database was examined for duplicate records, missing values, and data consistency. Variables with more than 20% missing values were initially considered for exclusion. Accordingly, micropore volume and pore diameter were excluded from subsequent analyses because of their high degree of missingness. Although pore volume also exceeded this threshold (32.2% missing), it was retained owing to its well-established role as a key structural descriptor governing CO₂ adsorption. Missing values for pore volume were imputed using the median value of the corresponding adsorbent family. BET surface area (6.3% missing) was similarly imputed using the family-median approach, whereas temperature, CO₂ partial pressure, and CO₂ adsorption capacity contained negligible or no missing values. Table 3 summarizes the missing-value distribution and preprocessing strategy adopted for the variables used in model development. The final dataset comprised BET surface area, pore volume,

adsorption temperature, CO₂ partial pressure, and adsorbent family as input variables, with CO₂ adsorption capacity used as the prediction target.

Table 3. Missing-value summary and preprocessing strategy for the variables used in model development.

Variable	Missing (n)	Missing (%)	Preprocessing strategy
BET Surface Area (m ² g ⁻¹)	25	6.3	Family-median imputation
Pore Volume (cm ³ g ⁻¹)	128	32.2	Family-median imputation
Pressure	2	0.5	No treatment required
Temperature (°C)	0	0.0	No treatment required
CO ₂ Partial Pressure (bar)	0	0.0	No treatment required
CO ₂ Adsorption Capacity (mmol g ⁻¹)	0	0.0	Target variable

2.3. Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) was conducted to examine the characteristics of the curated dataset prior to feature selection and machine learning model development. Descriptive statistical analysis was performed to summarize the central tendency and variability of the numerical variables. The distribution of adsorbent families and material categories was evaluated to assess dataset composition, while missing-value analysis was conducted to verify data completeness following preprocessing. Histograms and boxplots were used to examine variable distributions and identify potential outliers. Pearson correlation analysis was subsequently performed to investigate linear relationships among the numerical variables and to assess potential multicollinearity. The insights obtained from the EDA were used to support feature selection and guide the development of robust and interpretable machine learning models.

2.4. Feature Selection

Feature selection was performed using Recursive Feature Elimination with Cross-Validation (RFECV) to identify the optimal subset of predictors while reducing redundancy (Abdelbaky et al., 2021). Candidate variables included structural descriptors, operating conditions, and the categorical variable representing the adsorbent family. Variables with excessive missing values (micropore volume and pore diameter) were excluded during data preprocessing, whereas pore volume was retained owing to its recognized importance as a structural descriptor for CO₂ adsorption. Total pressure was excluded from the final feature set, while CO₂ partial pressure was retained as the thermodynamically relevant operating variable. The final predictor set consisted of

BET surface area, pore volume, adsorption temperature, CO₂ partial pressure, and adsorbent family. Although the contribution of the categorical variable was also evaluated using permutation importance and SHAP analyses, these were used primarily to improve model interpretability rather than for feature selection.

2.5. Machine Learning Models

Seven ensemble machine learning algorithms were evaluated for predicting CO₂ adsorption capacity, including Random Forest (RF)(Breiman, 2001), HistGradient Boosting (HGB), Gradient Boosting (GB)(Friedman), Extra Trees (ET)(Geurts et al.), Extreme Gradient Boosting (XGBoost)(Chen and Guestrin), Light Gradient Boosting Machine (LightGBM)(Ke et al., 2017), and CatBoost(Prokhorenkova et al., 2018). Model development and evaluation were performed using Python within the Google Colab environment. The curated dataset was trained using repeated cross-validation to ensure robust performance assessment. Model performance was evaluated using the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). Two predictive scenarios were investigated: Model I, incorporating adsorption data collected under all operating pressures, and Model II, comprising only ambient-pressure adsorption data to assess the influence of operating conditions on predictive performance.

2.6. Explainable Artificial Intelligence

To enhance the interpretability of the machine learning models, explainable artificial intelligence (XAI) techniques were employed to quantify the influence of individual predictors on CO₂ adsorption capacity. SHapley Additive exPlanations (SHAP) were used to evaluate both global feature importance and local feature contributions, providing a consistent interpretation of model predictions(Lundberg and Lee, 2017). Partial Dependence Plots (PDPs) were generated to visualize the marginal effect of key numerical variables on the predicted adsorption capacity while averaging over the remaining features (Fisher et al., 2019). In addition, permutation importance analysis was performed to assess the relative contribution of each predictor to model performance by measuring the reduction in predictive accuracy following random feature permutation(Friedman). Collectively, these complementary XAI techniques were used to improve model transparency and facilitate the interpretation of structure–property relationships governing CO₂ adsorption.

3. Results and Discussion

3.1. Characteristics of the Curated Database

The final curated database comprised 398 experimentally reported CO₂ adsorption measurements extracted from 67 peer-reviewed publications, representing four major adsorbent families: Carbon-based Adsorbents, Organic Porous Materials, Inorganic Porous Adsorbents, and Metal Oxide Adsorbents. As shown in Figure 2, Organic Porous Materials represented the largest proportion of the dataset (37.9%), followed by Metal Oxide Adsorbents (23.6%), Carbon-based Adsorbents

(23.1%), and Inorganic Porous Adsorbents (15.3%). The relatively balanced distribution among the four families provides a reliable basis for comparative machine learning analysis. The descriptive statistics of the numerical variables are presented in Table 4. The database covers a broad range of structural properties and operating conditions, with BET surface area ranging from 2.4 to 4023 m² g⁻¹, pore volume from 0.01 to 2.62 cm³ g⁻¹, adsorption temperature from 0 to 750 °C, and CO₂ partial pressure from 0.05 to 200 bar. Hence, CO₂ adsorption capacity spans from 0 to 144 mmol g⁻¹, reflecting the substantial diversity of materials and experimental conditions reported in the literature. This wide variability provides a robust foundation for developing machine learning models capable of identifying structure–property relationships across different adsorbent families and operating conditions.

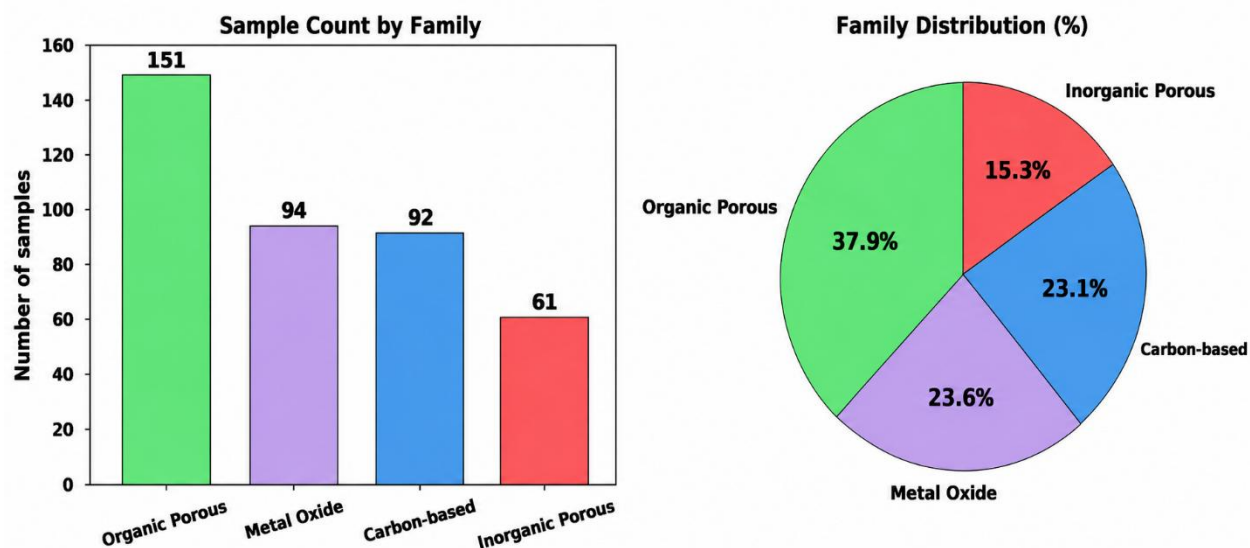


Figure 2. Sample count and percentage contribution of each adsorbent family (n = 398).

Table 4. Descriptive statistics of the numerical variables (n = 398)

Variable	n	Min	Q1	Median	Mean	Q3	Max	SD
BET Surface Area [m ² /g]	373	2.4	260	630	863.92	1043	4023	858.99
Pore Volume [cc/g]	270	0.01	0.25	0.45	0.61	0.82	2.62	0.54
Pressure [bar]	396	0.1	1	1	11.67	1	200	41.85
Temperature [°C]	398	0	15	25	67.9	50	750	137.56
CO ₂ Partial Pressure [bar]	398	0.05	1	1	11.53	1	200	41.77

CO₂ Adsorption Capacity [mmol g⁻¹]	398	0	1.02	1.77	6.88	3.45	144	20.92
---	-----	---	------	------	------	------	-----	-------

3.2. Exploratory Data Analysis

3.2.1. Distribution of Numerical Variables

The distribution of the numerical variables included in the curated database is illustrated in Figure 3, while the corresponding skewness and kurtosis values are summarized in Table 5. All variables exhibited positively skewed distributions, indicating that most observations were concentrated at relatively low values with a limited number of high-value measurements. This trend was particularly evident for CO₂ adsorption capacity (skewness = 4.75), CO₂ partial pressure (4.17), and temperature (3.09), reflecting the predominance of adsorption experiments conducted under ambient or near-ambient conditions and the comparatively smaller number of studies performed at elevated temperatures or high-pressure conditions. Similarly, all variables showed positive excess kurtosis, confirming the presence of heavy-tailed distributions. The highest kurtosis values were observed for CO₂ partial pressure (15.86), pressure (15.76), and CO₂ adsorption capacity (22.41), suggesting that a small number of extreme observations contribute substantially to the overall data variability. These observations are consistent with the heterogeneous nature of literature-derived adsorption datasets, which combine a wide range of materials and experimental conditions (Hajiali et al., 2025; Yuan et al., 2021). The observed non-normality indicates that conventional linear statistical assumptions are not fully satisfied. However, these characteristics do not compromise the subsequent machine learning analysis because ensemble tree-based algorithms are inherently robust to skewed distributions and non-linear relationships. Consequently, the diverse distribution of the variables provides a suitable basis for investigating complex structure–property relationships governing CO₂ adsorption across different adsorbent families (Chowdhury et al.).

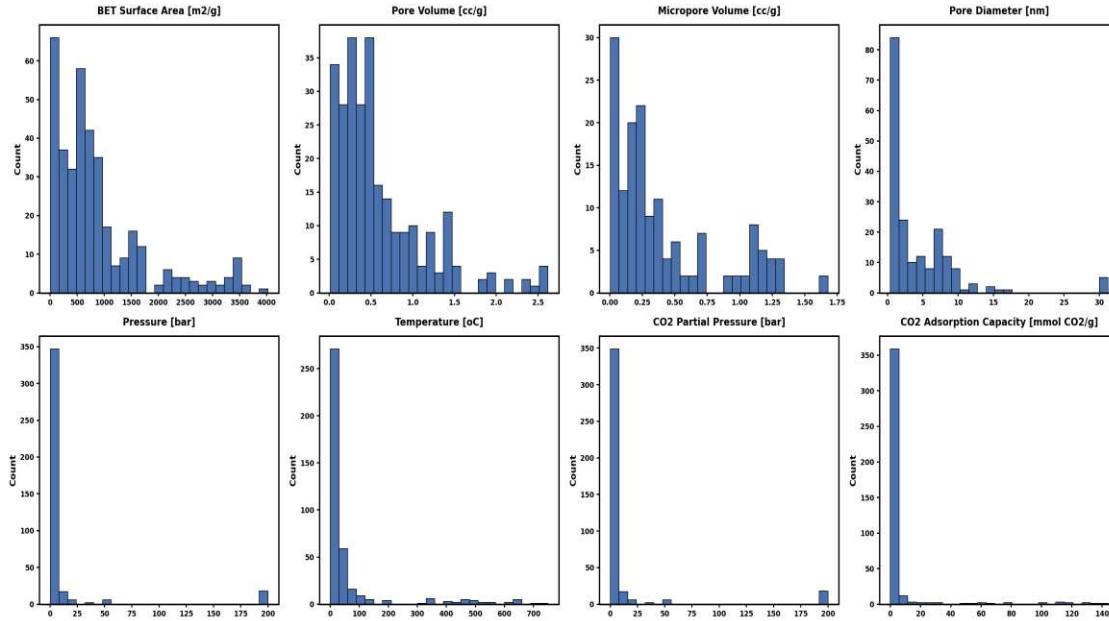


Figure 3. Histograms of the six numerical variables.

Table 5. Skewness and excess kurtosis of the numerical variables included in the exploratory data analysis.

Variable	Skewness	Excess Kurtosis
BET Surface Area	1.62	2.21
Pore Volume	1.65	2.69
Pressure	4.16	15.76
Temperature	3.09	8.76
CO ₂ Partial Pressure	4.17	15.86
CO ₂ Adsorption Capacity	4.75	22.41

3.2.2. Correlation Analysis and Multicollinearity

Pearson and Spearman correlation analyses were performed to examine the relationships among the numerical variables included in the final machine learning dataset (Figure 4). Pearson correlation evaluates linear associations, whereas Spearman correlation assesses monotonic relationships and is less sensitive to non-normal data distributions and extreme observations. The combined use of both methods provides a comprehensive assessment of the relationships among the predictor variables. The correlation matrices revealed an almost perfect positive correlation between total pressure and CO₂ partial pressure, indicating that the two variables convey nearly

identical information within the curated database. This strong relationship arises because the majority of the collected adsorption experiments were conducted using pure or nearly pure CO₂ streams, where the total system pressure closely approximates the CO₂ partial pressure. Although a limited number of studies, primarily involving flue-gas or dilute CO₂ conditions, exhibited differences between these two variables, the overall correlation remained sufficiently high to indicate severe redundancy. Because CO₂ partial pressure represents the thermodynamic driving force governing adsorption equilibrium, it was retained as the pressure-related descriptor in the predictive models, whereas total pressure was excluded to eliminate redundant information while preserving the most physically meaningful variable for adsorption modelling (Ghaedi, 2021; Söğüt and Gülcan, 2023). To further evaluate multicollinearity among the candidate predictors, Variance Inflation Factor (VIF) analysis was performed (Table 6). BET surface area (VIF = 4.97), pore volume (VIF = 4.86), and temperature (VIF = 1.49) all exhibited acceptable VIF values, indicating that these variables could be retained without introducing problematic multicollinearity. In contrast, total pressure and CO₂ partial pressure produced extremely high VIF values (> 67,000), confirming their near-perfect multicollinearity (Perera et al., 2012). Consequently, total pressure was removed from the final predictor set, while CO₂ partial pressure was retained as the physically relevant descriptor governing adsorption equilibrium. Overall, the correlation and multicollinearity analyses confirmed that the selected predictor variables provide complementary rather than redundant information for machine learning modelling. Eliminating total pressure improved the numerical stability of the predictive models while preserving the descriptor with the greatest physicochemical significance. Importantly, this preprocessing step also established the basis for one of the principal findings of this study, namely the dominant influence of CO₂ partial pressure on adsorption prediction observed in the subsequent machine learning and explainable AI analyses (O'brien, 2007) (Ojong et al., 2026). Overall, the identification and removal of redundant variables before model development reduced numerical instability while preserving the physicochemically meaningful adsorption descriptors, thereby improving both model robustness and the interpretability of the subsequent explainable AI analysis.

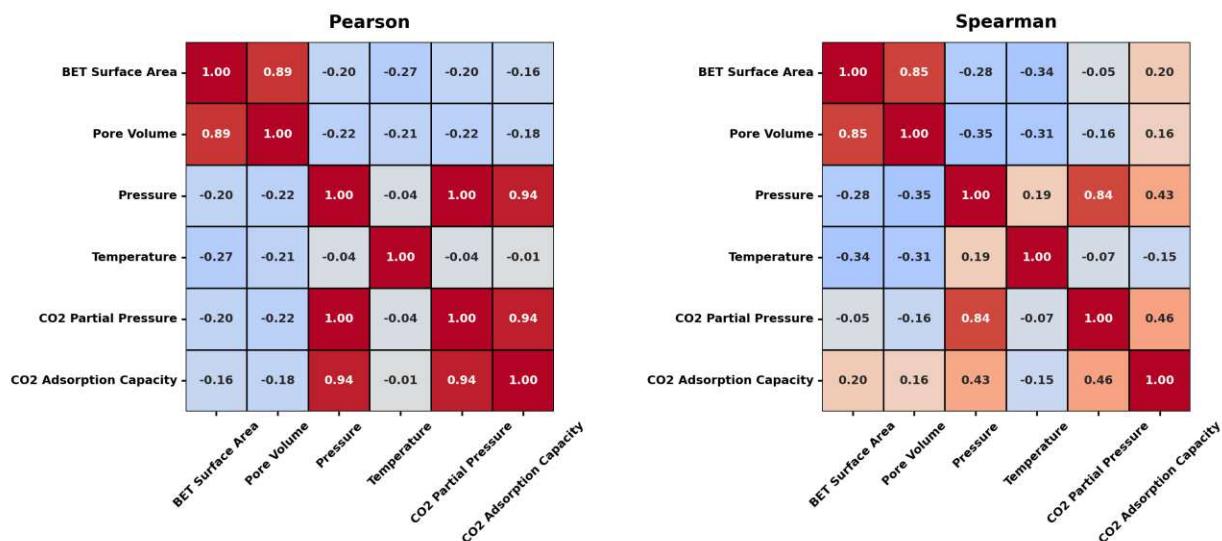


Figure 4. Pearson (linear) and Spearman (rank/monotonic) correlation matrices for the five practical-model variables.

Table 6. Variance Inflation Factor (VIF) analysis of the candidate predictor variables

Variable	VIF
BET Surface Area	4.97
Pore Volume	4.86
Pressure	67,702.98 (VIF ↑)
Temperature	1.49
CO ₂ Partial Pressure	67,688.40 (VIF ↑)

3.2.3. Principal Component Analysis

Principal Component Analysis (PCA) was performed using the four numerical predictors retained in the final feature set to visualize the multivariate distribution of the adsorbent families (Jolliffe and Cadima, 2016). The PCA score plot (Figure 5) shows that the first two principal components account for **75.9%** of the total variance (PC1 = 49.7%, PC2 = 26.2%), indicating that the principal trends in the dataset are well represented in the two-dimensional projection. Although certain family-specific tendencies are evident, considerable overlap is observed among the four adsorbent families, particularly near the origin of the score plot. This overlap indicates that the selected structural descriptors alone are insufficient to clearly separate the material families. Instead, the most distinct groupings correspond to a limited number of samples measured under extreme operating conditions. Organic Porous Materials are primarily separated toward positive PC2,

reflecting experiments conducted at elevated CO₂ partial pressures, whereas Metal Oxide Adsorbents extend toward negative PC2 because they were predominantly evaluated at higher adsorption temperatures. These observations suggest that operating conditions contribute more strongly to the observed separation than intrinsic material chemistry (Namdeo et al., 2023). Overall, the PCA demonstrates that the curated database is highly heterogeneous and that simple family-based classification is insufficient to explain CO₂ adsorption performance. This finding provides additional support for the use of explainable machine learning to disentangle the relative contributions of operating conditions and material descriptors (Sung et al.).

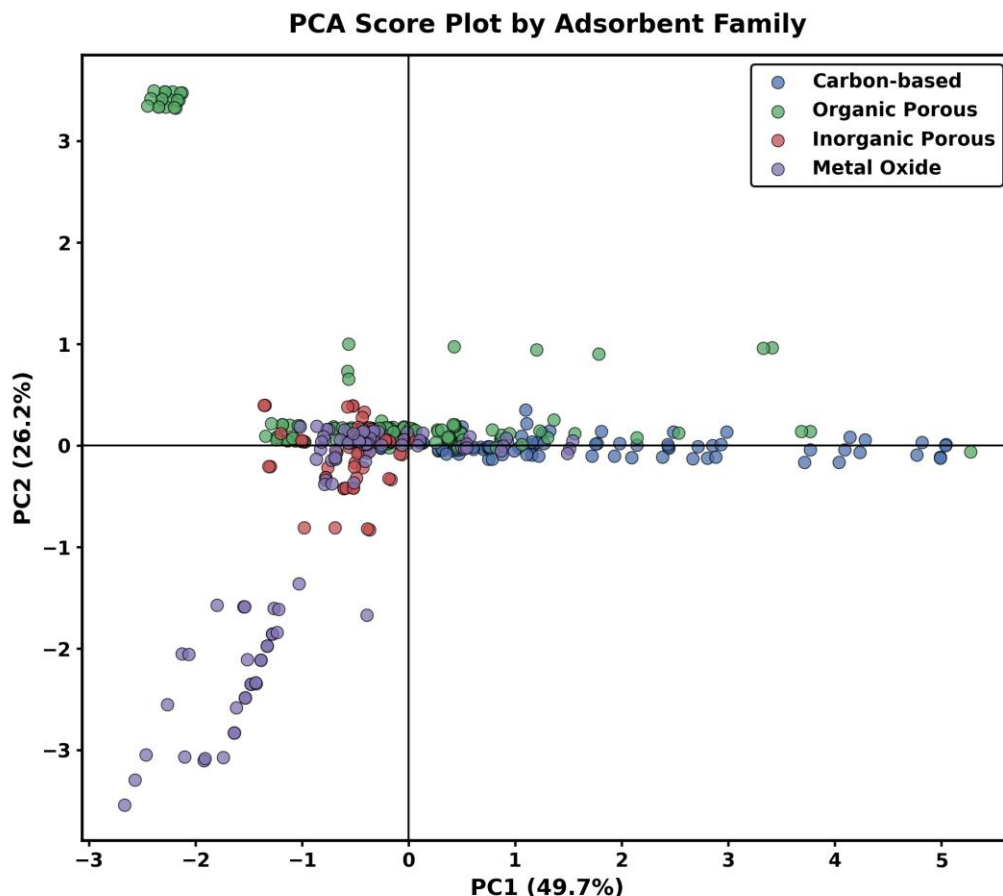


Figure 5. PCA score plot (PC1 vs PC2) coloured by adsorbent family, based on BET Surface Area, Pore Volume, Temperature, and CO₂ Partial Pressure.

3.3. Family-Wise Comparison of CO₂ Adsorption Performance

3.3.1 Comparison of Adsorption Performance Across Adsorbent Families

The adsorption performance of the four adsorbent families was first compared using the complete curated database to identify general differences in structural properties and CO₂ uptake. As

summarized in Table 7, Organic Porous Materials exhibited the highest average CO₂ adsorption capacity (14.32 mmol g⁻¹), despite having a lower mean BET surface area (746.68 m² g⁻¹) than Carbon-based Adsorbents (1716.29 m² g⁻¹). In contrast, Carbon-based Adsorbents, Metal Oxide Adsorbents, and Inorganic Porous Adsorbents showed substantially lower average adsorption capacities of 2.81, 2.28, and 1.67 mmol g⁻¹, respectively. These observations indicate that adsorption capacity cannot be explained solely by BET surface area and suggest that operating conditions and other physicochemical factors also play important roles (Mehrmohammadi and Ghaemi, 2023). The distributions of CO₂ adsorption capacity, BET surface area, and pore volume for each adsorbent family are illustrated in Figure 6. Considerable variability is evident within each family, together with numerous high-value observations, particularly for Organic Porous Materials. The broad interquartile ranges and the presence of outliers highlight the heterogeneous nature of the compiled database and further demonstrate that comparisons based solely on average values may be misleading (Sekizkardes et al., 2022). These findings demonstrate that direct comparison of adsorption capacities across adsorbent families without accounting for operating pressure may lead to misleading conclusions regarding intrinsic material performance (Shi et al., 2024). Consequently, subsequent machine learning analyses were designed to distinguish the effects of operating conditions from intrinsic material descriptors.

3.3.2 Influence of Operating Pressure on Family Comparison

To investigate whether the apparent superiority of Organic Porous Materials was influenced by operating conditions, the adsorption capacities were further separated according to pressure regime (Figure 7). A substantial proportion of the exceptionally high adsorption capacities for Organic Porous Materials originated from measurements conducted above 1 bar, whereas Metal Oxide Adsorbents were evaluated exclusively under ambient-pressure conditions. When only ambient-pressure data were considered, the differences among the four families became markedly smaller, with Carbon-based Adsorbents exhibiting the highest mean adsorption capacity (2.69 mmol g⁻¹), followed by Metal Oxide Adsorbents (2.28 mmol g⁻¹), Organic Porous Materials (2.12 mmol g⁻¹), and Inorganic Porous Adsorbents (1.25 mmol g⁻¹), as summarized in Table 8. The statistical significance of these differences was assessed using the Shapiro–Wilk (Shapiro and Wilk, 1965), Levene's (Levene, 1960), and Kruskal–Wallis (Kruskal and Wallis, 1952) tests (Table 9). The results confirmed that the adsorption capacities differed significantly among the four adsorbent families ($p < 0.05$). However, the pressure-stratified analysis demonstrates that part of the apparent performance advantage observed in the complete dataset arises from differences in testing conditions rather than intrinsic material characteristics (Ghawche et al., 2026). This finding provides the principal motivation for the subsequent machine learning analysis, which aims to separate the influence of operating conditions from the intrinsic properties of the adsorbent materials (Mehrmohammadi and Ghaemi, 2023).

Table 7. Descriptive statistics of the four adsorbent families included in the curated database.

Family	n	% of Total	Mean Capacity (mmol g ⁻¹)	Mean BET (m ² g ⁻¹)
Organic Porous Materials	151	37.9%	14.32	746.68
Metal Oxide Adsorbents	94	23.6%	2.28	286.85
Carbon-based Adsorbents	92	23.1%	2.81	1716.29
Inorganic Porous Adsorbents	61	15.3%	1.67	607.12

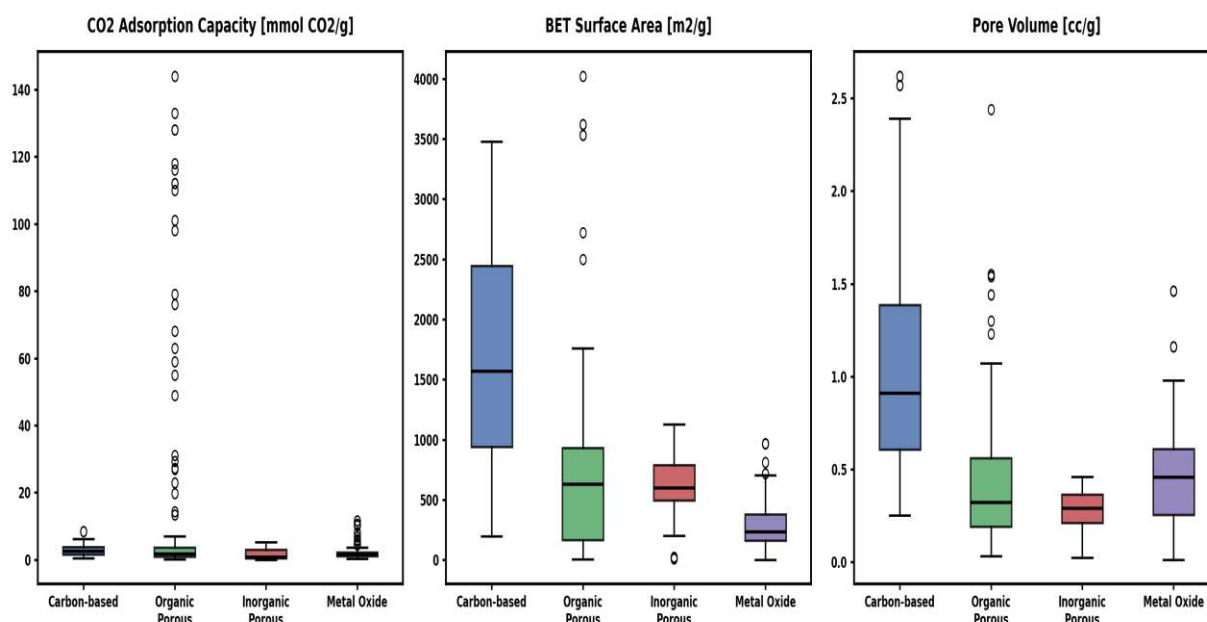


Figure 6. Boxplots of CO₂ Adsorption Capacity, BET Surface Area, and Pore Volume across the four adsorbent families.

Table 8. Mean and median CO₂ adsorption capacity of the four adsorbent families under ambient-pressure conditions.

Family	n (ambient)	Mean Capacity (mmol g ⁻¹)	Median Capacity (mmol g ⁻¹)
Carbon-based Adsorbents	85	2.69	2.51
Metal Oxide Adsorbents	94	2.28	1.51
Organic Porous Materials	111	2.12	1.45

Inorganic Porous Adsorbents	48	1.25	0.62
-----------------------------	----	------	------

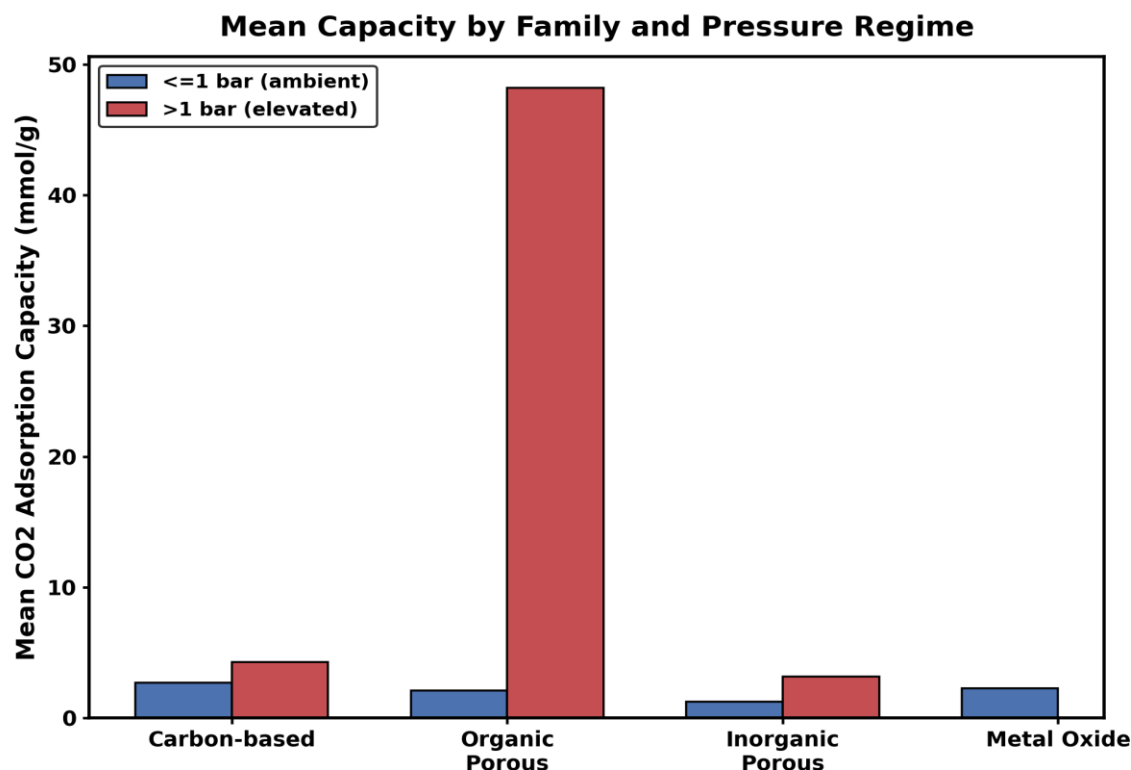


Figure 7. Mean CO₂ adsorption capacity by family, split by pressure regime (≤ 1 bar vs > 1 bar). Metal Oxide has no elevated-pressure bar because it was never tested above 1 bar.

The statistical significance of the differences in CO₂ adsorption capacity among the four adsorbent families was evaluated using the Shapiro–Wilk, Levene's, and Kruskal–Wallis tests (Table 9). The Shapiro–Wilk test indicated that adsorption capacity was not normally distributed within any of the four families (all $p < 0.0001$), while Levene's test confirmed significant heterogeneity of variances ($p = 0.0001$). Because the assumptions of normality and homogeneity of variance required for parametric analysis were not satisfied, the non-parametric Kruskal–Wallis test was employed. The Kruskal–Wallis test revealed significant differences in CO₂ adsorption capacity among the four adsorbent families for both the complete dataset ($H = 28.21$, $p = 3.3 \times 10^{-6}$) and the ambient-pressure subset ($H = 41.67$, $p = 4.7 \times 10^{-9}$). Interestingly, the statistical significance increased after restricting the analysis to ambient-pressure measurements, indicating that the observed differences are not solely attributable to the imbalance in testing pressure across the literature dataset. These findings demonstrate that although operating pressure influences the apparent adsorption performance of different material families, statistically significant differences

remain after controlling for this experimental factor. Consequently, both intrinsic material characteristics and operating conditions contribute to the observed adsorption behaviour, providing the basis for the subsequent machine learning and explainable AI analyses (Nait Amar et al.; Namdeo et al., 2023). To complement the p-values reported above, effect sizes were estimated using the eta-squared approximation for the Kruskal–Wallis statistic, $\eta^2 = (H - k + 1)/(n - k)$, where k is the number of families. This yielded $\eta^2 = 0.064$ for the complete dataset and $\eta^2 = 0.116$ for the ambient-pressure subset, corresponding to a medium and a medium-to-large effect, respectively, by conventional benchmarks. These values indicate that family membership explains a modest but non-trivial share of the variance in adsorption capacity, and that this share increases once the pressure confound is removed, reinforcing the interpretation above.

Table 9. Statistical evaluation of family-wise differences in CO₂ adsorption capacity under the complete dataset and ambient-pressure conditions.

Test	Statistic	p-value	Conclusion
Shapiro-Wilk (per family)	W = 0.32 – 0.93	all p < 0.0001	Capacity is non-normal in every family
Levene's test	F = 7.09	p = 0.0001	Variances are unequal across families
Kruskal-Wallis (full, n=398)	H = 28.21	p = 3.3×10^{-6}	Capacities differ significantly across families
Kruskal-Wallis (ambient only, n=338)	H = 41.67	p = 4.7×10^{-9}	Difference holds — in fact strengthens — after removing the pressure confound

3.4 Validation of the Curated Dataset

3.4.1 Pore Volume Sensitivity Analysis

Because pore volume contained the highest proportion of missing values among the retained predictors (32.2%), a sensitivity analysis was performed to evaluate its contribution to model performance (Little and Rubin, 2019). A Random Forest model with five-fold cross-validation was trained using two scenarios (Breiman, 2001): (i) the complete curated dataset after family-median imputation of missing pore volume values (n = 398), and (ii) the subset containing only experimentally measured pore volume values (n = 270). The predictive performance with and without pore volume is presented in Figure 8. As shown in Figure 8, including pore volume resulted in a modest improvement in predictive performance for the complete dataset, increasing the mean cross-validated R² from 0.943 ± 0.044 to 0.957 ± 0.027 . This limited improvement is expected because approximately one-third of the pore volume values were imputed using family-median values, thereby reducing the amount of unique structural information contributed by this descriptor. A more pronounced improvement was observed when the analysis was restricted to samples with experimentally measured pore volume values. Under these conditions, the inclusion

of pore volume increased the mean R^2 from 0.889 ± 0.119 to 0.953 ± 0.044 , while simultaneously reducing the variability of the cross-validation results. These findings indicate that experimentally measured pore volume provides valuable structural information for predicting CO_2 adsorption capacity and improves model robustness. Overall, the sensitivity analysis demonstrates that retaining pore volume is justified despite its relatively high proportion of missing values (Zhu et al., 2020). Although family-median imputation inevitably reduces its predictive contribution, the inclusion of pore volume consistently improves model performance compared with its exclusion (Iddrisu et al., 2025). Accordingly, pore volume was retained as one of the structural descriptors for the subsequent machine learning analyses.

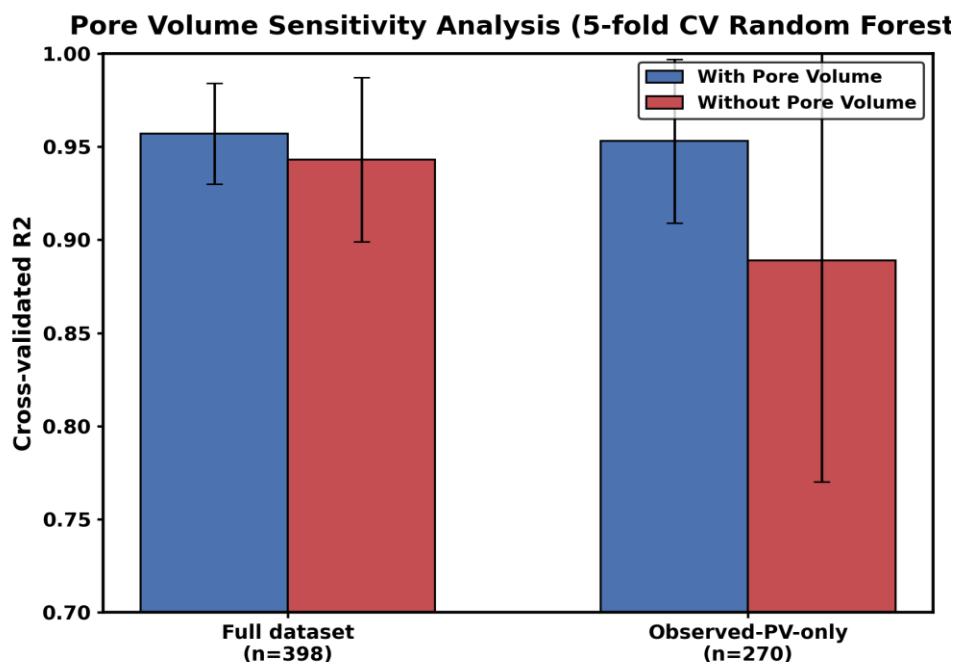


Figure 8. Effect of pore volume inclusion on Random Forest predictive performance under five-fold cross-validation for the complete curated dataset ($n = 398$) and the subset containing experimentally measured pore volume values ($n = 270$).

3.5 Machine learning model performance

3.5.1 Comparative Evaluation of Machine Learning Algorithms

To identify the most suitable predictive model for CO_2 adsorption capacity, seven ensemble machine learning algorithms, namely Random Forest (RF), Gradient Boosting (GB), HistGradientBoosting (HGB), Extra Trees (ET), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and CatBoost, were evaluated using repeated cross-validation. Model performance was assessed using the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). The comparison was performed for two independent datasets: Model I, comprising the complete dataset with all pressure conditions, and

Model II, containing only ambient-pressure (≤ 1 bar) observations. The predictive performances of all algorithms are summarized in Table 10. For Model I, all tree-based ensemble algorithms achieved high predictive accuracy, with Gradient Boosting providing the highest performance ($R^2 = 0.930$), closely followed by XGBoost ($R^2 = 0.929$), CatBoost ($R^2 = 0.921$), and Random Forest ($R^2 = 0.920$). In contrast, LightGBM and HistGradientBoosting exhibited noticeably lower predictive accuracy, with R^2 values of 0.824 and 0.782, respectively. The consistently high performance of the leading algorithms indicates that the relationship between the selected descriptors and CO₂ adsorption capacity can be effectively captured by ensemble learning methods when the full range of operating pressures is considered. A markedly different trend was observed for Model II, where the analysis was restricted to ambient-pressure conditions. Under these conditions, predictive performance decreased for all algorithms, with the highest performance obtained by HistGradientBoosting ($R^2 = 0.417$), followed by CatBoost ($R^2 = 0.381$) and LightGBM ($R^2 = 0.357$). Conventional ensemble methods, including Random Forest, Gradient Boosting, and XGBoost, exhibited only moderate predictive capability, while Extra Trees showed the lowest performance ($R^2 = 0.149$). The substantial reduction in predictive accuracy from Model I to Model II highlights the dominant influence of CO₂ partial pressure on adsorption capacity prediction (Zhu et al., 2020). When high-pressure experiments are included, the machine learning algorithms can readily learn the strong pressure–capacity relationship, resulting in high predictive performance. However, once the pressure range is restricted to ambient conditions, this dominant source of variation is largely removed, and the prediction becomes increasingly dependent on intrinsic material properties, including surface area, pore structure, and adsorbent family (Abdi et al., 2021). Consequently, the lower R^2 values observed for Model II should not be interpreted as model failure; rather, they reflect the inherently greater complexity of predicting adsorption behaviour when operating pressure is no longer the primary driving factor. As a result, the comparison demonstrates that ensemble learning methods provide robust predictive capability for CO₂ adsorption capacity across diverse adsorbent materials. Furthermore, the contrasting performances between the two datasets provide additional evidence that operating pressure strongly influences the apparent predictive relationships within the compiled literature database (Iyiola et al., 2025), supporting the statistical observations presented in Section 3.3. Overall, the large performance difference between Model I and Model II highlights the dominant contribution of operating pressure to the predictive accuracy of literature-derived adsorption datasets. When pressure variation is removed, adsorption prediction becomes substantially more challenging because the remaining structural descriptors explain a considerably smaller proportion of the observed variability.

Table 10. Performance comparison of seven ensemble machine learning algorithms for predicting CO₂ adsorption capacity using Model I (all-pressure dataset) and Model II (ambient-pressure dataset).

Dataset	Model	R^2	RMSE (mmol/g)	MAE (mmol/g)
---------	-------	-------	------------------	-----------------

Model I (all-pressure)	GradientBoosting	0.930	3.567	1.496
Model I (all-pressure)	XGBoost	0.929	3.574	1.386
Model I (all-pressure)	CatBoost	0.921	4.002	1.556
Model I (all-pressure)	Random Forest	0.920	4.058	1.527
Model I (all-pressure)	ExtraTrees	0.905	4.303	1.584
Model I (all-pressure)	LightGBM	0.824	7.085	3.305
Model I (all-pressure)	HistGradientBoosting	0.782	8.224	3.734
Model II (ambient-only)	HistGradientBoosting	0.417	1.698	0.788
Model II (ambient-only)	CatBoost	0.381	1.699	0.722
Model II (ambient-only)	LightGBM	0.357	1.764	0.798
Model II (ambient-only)	Random Forest	0.344	1.732	0.776
Model II (ambient-only)	GradientBoosting	0.342	1.730	0.783
Model II (ambient-only)	XGBoost	0.287	1.770	0.759
Model II (ambient-only)	ExtraTrees	0.149	1.905	0.806

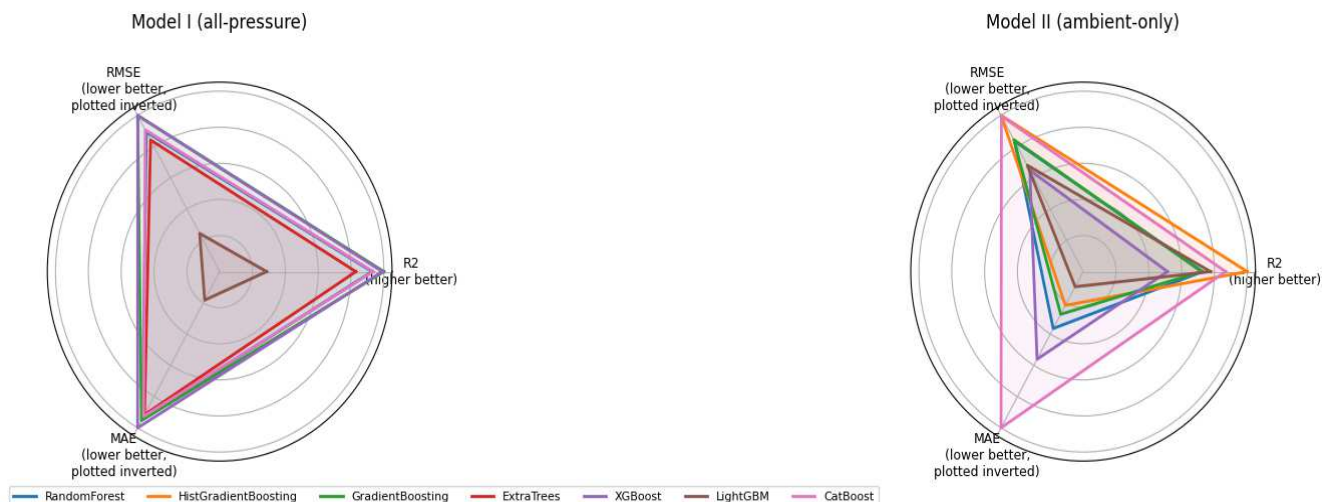


Figure 9. Multi-metric radar comparison (R^2 , RMSE, MAE) across all seven algorithms.

3.5.2 Selection of the Final Prediction Model

Although several ensemble learning algorithms achieved competitive predictive performance, model selection for subsequent interpretation considered both predictive accuracy and model interpretability. Gradient Boosting achieved the highest predictive performance for Model I, whereas HistGradientBoosting provided the best results for Model II under ambient-pressure conditions (Table 10). The differences among the top-performing algorithms, however, were relatively small, indicating that all leading ensemble methods were capable of capturing the complex nonlinear relationships governing CO_2 adsorption. To investigate the physicochemical significance of the selected descriptors rather than solely maximize predictive accuracy, the subsequent explainable artificial intelligence (XAI) analysis was conducted using the Random Forest model. Random Forest offers robust predictive performance while providing well-established frameworks for feature importance analysis, permutation importance, and SHAP interpretation, making it particularly suitable for identifying the relative contributions of structural descriptors and operating conditions to adsorption capacity (Lundberg and Lee, 2017; Salih et al.). Although Gradient Boosting achieved marginally higher predictive accuracy, Random Forest was selected for the explainable AI analysis because its bagging-based ensemble structure produces more stable SHAP value estimates across cross-validation folds and is less sensitive to the sequential error-correction behavior of boosting algorithms, which can amplify the influence of dominant features such as CO_2 partial pressure. This choice prioritizes interpretive stability over the marginal accuracy gain of boosting-based models. The comparison between Model I and Model II further demonstrated that the predictive performance of all algorithms decreased substantially after restricting the analysis to ambient-pressure conditions. This consistent trend confirms that the reduction in prediction accuracy is primarily associated with the removal of pressure-dependent information rather than the choice of machine learning algorithm. Consequently, the explainable AI analyses presented in the following sections focus on identifying

how the relative importance of structural descriptors changes once the dominant influence of operating pressure is reduced (Rudin, 2019).

3.6 Explainable Machine Learning Analysis

3.6.1 Global SHAP Analysis

To interpret the predictions generated by the machine learning models, SHapley Additive exPlanations (SHAP) were employed to quantify the contribution of each predictor to the predicted CO₂ adsorption capacity. Unlike conventional feature importance measures, SHAP provides both the magnitude and direction of each variable's influence on individual predictions, enabling a comprehensive interpretation of the relationships learned by the models (Lundberg and Lee, 2017). Global SHAP summary plots for Model I and Model II are presented in Figure 10. For Model I (all-pressure dataset) Fig. 10 (a), CO₂ partial pressure overwhelmingly dominated the prediction process, exhibiting substantially larger SHAP values than all other predictors. Samples with high CO₂ partial pressure consistently produced positive SHAP values, indicating a strong positive contribution to the predicted adsorption capacity, whereas low-pressure conditions contributed negatively. This finding confirms the results of the multicollinearity analysis (Section 3.2), the family-wise comparison (Section 3.3), and the machine learning performance evaluation (Section 3.5), all of which identified operating pressure as the principal factor governing adsorption behavior within the compiled database. After restricting the analysis to ambient-pressure conditions (Model II) Fig. 10 (b), the global SHAP distribution changed markedly. The influence of CO₂ partial pressure decreased considerably, while BET surface area, pore volume, temperature, and Main Family exhibited substantially larger SHAP contributions. This shift indicates that, once the dominant pressure effect is controlled, the model increasingly relies on intrinsic material characteristics rather than operating conditions to predict adsorption capacity. The comparison between Models I and II demonstrates that the apparent importance of predictor variables depends strongly on the operating conditions represented within the dataset. Under the full pressure range, machine learning models primarily learn the pressure–capacity relationship, which masks the influence of structural descriptors (Kou et al., 2025). In contrast, restricting the analysis to ambient-pressure measurements reveals the contribution of physicochemical properties that are directly related to adsorption behaviour. Consequently, the ambient-pressure model provides a more meaningful representation of structure–property relationships despite its lower predictive accuracy. Overall, the global SHAP analysis demonstrates that explainable artificial intelligence not only identifies the variables responsible for high predictive performance but also distinguishes between predictors associated with experimental conditions and those reflecting intrinsic adsorbent properties. These global observations are further explored in the following section through family-specific SHAP analysis (Longe et al., 2025). Although CO₂ partial pressure dominated the global feature importance analysis, this finding should not be interpreted as evidence that structural descriptors are unimportant. Rather, the wide pressure range represented in the curated database masks a substantial proportion of the structure–property relationships,

which become increasingly evident under restricted operating conditions, as demonstrated by the ambient-pressure analyses(Li et al., 2023).

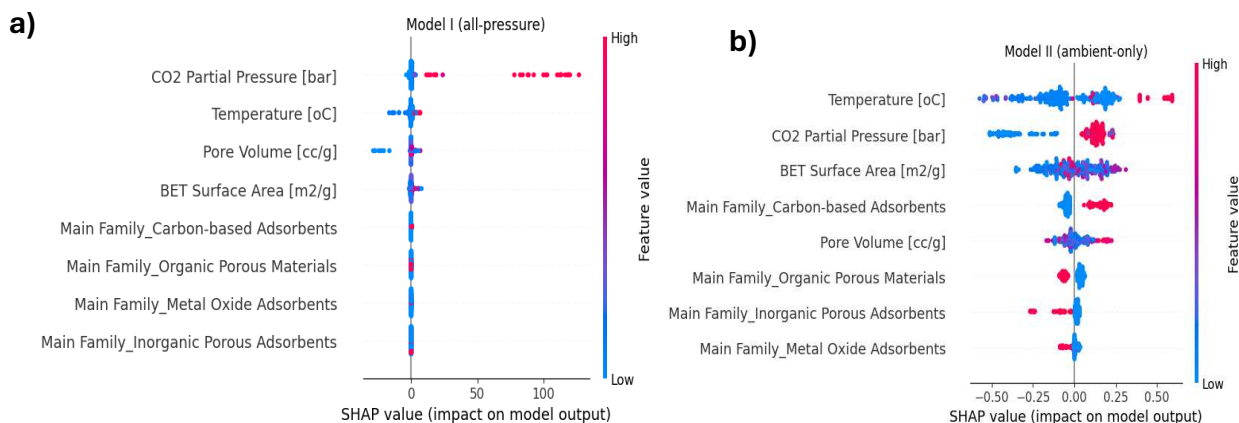


Figure 10 Global SHAP summary plot for (a) Model I (all-pressure dataset) (b) Model II (ambient-only).

3.6.2 Family-Specific SHAP Analysis

To further investigate the influence of operating conditions and material chemistry, the SHAP analysis was performed separately for each of the four adsorbent families. The mean absolute SHAP ($|\text{SHAP}|$) values for each predictor are summarized in Table 11, while the corresponding visualizations are presented in Figures 13 (a) and 13 (b) for Model I and Model II, respectively. For Model I (all-pressure dataset), pronounced differences were observed among the adsorbent families. The most striking result was obtained for Organic Porous Materials, where CO₂ partial pressure exhibited a mean $|\text{SHAP}|$ value of 22.92, whereas the corresponding values for Carbon-based, Inorganic Porous, and Metal Oxide adsorbents ranged only from 0.44 to 0.55 (Table 11). This exceptionally large difference demonstrates that the predictions for Organic Porous Materials were dominated by operating pressure rather than intrinsic structural characteristics. Because this family contained most of the high-pressure adsorption experiments in the compiled database, the machine learning model primarily associated high adsorption capacities with elevated CO₂ partial pressures instead of the material properties themselves. Consequently, the apparent superiority of Organic Porous Materials observed in the full dataset is largely attributable to differences in experimental conditions rather than fundamental adsorption behavior. A markedly different pattern emerged for Model II, where the analysis was restricted to ambient-pressure measurements (Figure 11 (a)). Under these conditions, the SHAP values for all predictor variables fell within a much narrower range, and the extreme dominance of CO₂ partial pressure disappeared. Instead, temperature, BET surface area, and pore volume became consistently important predictors across the four adsorbent families, while the contribution of Main Family remained moderate but clearly

detectable. The substantially more balanced SHAP distributions indicate that, once the pressure confounding effect is removed, the machine learning model relies primarily on intrinsic material characteristics rather than experimental operating conditions. These observations provide direct quantitative evidence for the pressure–family confounding identified earlier in Sections 3.3 and 3.4. The family-specific SHAP analysis demonstrates that the dominant role of CO₂ partial pressure in Model I originates from the unequal distribution of operating pressures among the different adsorbent families rather than from universally stronger pressure sensitivity of one material class (Oviedo et al., 2022). After controlling for pressure, the predictor contributions become considerably more comparable across all families, indicating that the observed differences in adsorption behaviour more closely reflect genuine structure–property relationships. Overall, the family-specific SHAP analysis highlights the value of explainable artificial intelligence for distinguishing experimental bias from intrinsic material behaviour (Barnes et al., 2018). Beyond improving model transparency, this approach provides mechanistic insight into how different adsorbent families respond to structural descriptors and operating conditions, thereby supporting more reliable interpretation of machine learning models in adsorption research.

Table 11. Mean absolute SHAP ($|SHAP|$) values of the predictor variables for each adsorbent family in Model I (all-pressure dataset).

Feature	Carbon-based	Inorganic Porous	Metal Oxide	Organic Porous
BET Surface Area	0.36	0.20	0.17	0.65
CO ₂ Partial Pressure	0.44	0.47	0.55	22.92
Pore Volume	0.15	0.06	0.03	2.27
Temperature	0.45	0.47	1.14	1.58

Model I (all-pressure), mean $|SHAP|$ by family.

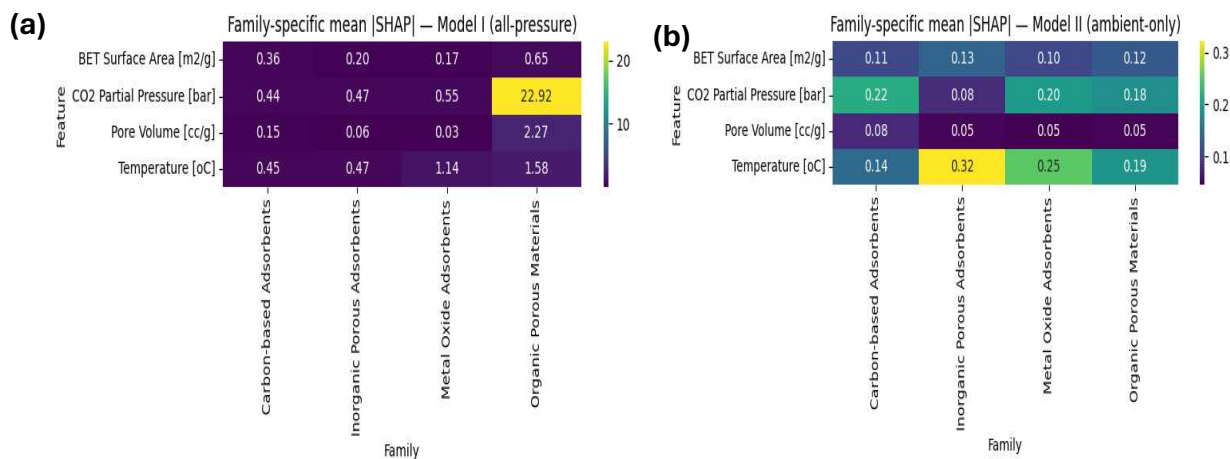


Figure 11. Family-specific SHAP importance for (a) Model I (all-pressure dataset) and (b) Model II (ambient-pressure dataset).

3.6.3 Partial Dependence Analysis of Temperature

To further interpret the relationship between adsorption temperature and CO₂ adsorption capacity, a partial dependence plot (PDP) was generated using the optimized Model II (ambient-pressure dataset). Unlike SHAP, which explains individual predictions, partial dependence analysis illustrates the average effect of a single predictor on the model response while averaging over all remaining variables. The resulting PDP for temperature is presented in Figure 12. The temperature response exhibits a distinct non-linear trend. At low temperatures, the predicted adsorption capacity decreases gradually with increasing temperature, reaching a minimum within the intermediate temperature range. Beyond approximately 350–400 °C, however, the predicted adsorption capacity increases sharply and continues to rise at higher temperatures. This behaviour indicates that the machine learning model successfully identified two different adsorption regimes represented within the compiled database. The initial decrease in adsorption capacity is consistent with the behaviour of physisorption-dominated adsorbents, including carbon-based materials, organic porous materials, and inorganic porous adsorbents. In these materials, CO₂ adsorption is an exothermic process, and increasing temperature weakens the interactions between CO₂ molecules and the adsorbent surface, resulting in lower adsorption capacities (Al-Mamoori et al., 2025). Conversely, the pronounced increase observed at elevated temperatures corresponds to chemisorption-dominated metal oxide adsorbents, such as calcium oxide and magnesium oxide, which require higher temperatures to activate the carbonation reaction and therefore exhibit enhanced CO₂ uptake under these conditions (Zheng et al., 2021). Importantly, this characteristic temperature profile was learned directly from the experimental database without explicitly providing the model with information regarding adsorption mechanisms. The ability of the machine learning model to recover this well-established physicochemical behavior demonstrates that the explainable AI framework captures meaningful adsorption phenomena rather than merely identifying statistical correlations. Consequently, the partial dependence analysis provides independent evidence supporting the mechanistic validity of the developed prediction models (Nunes).

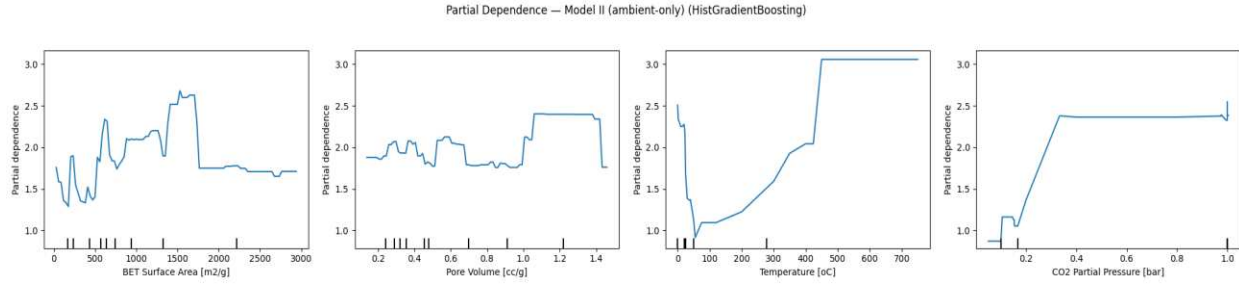


Figure 12 Partial dependence plot illustrating the effect of adsorption temperature on the predicted CO₂ adsorption capacity using Model II (ambient-pressure dataset).

3.7 Model Validation and Reliability Assessment

3.7.1 Predicted versus Experimental CO₂ Adsorption Capacity

To assess the agreement between the experimental observations and the machine learning predictions, the predicted CO₂ adsorption capacities were plotted against the experimentally measured values for the independent test datasets of Model I and Model II (Figure 13). The dashed diagonal line represents perfect agreement between measured and predicted values (Chicco et al., 2021; James, 2013). Model I exhibited excellent agreement, with most observations distributed closely around the 1:1 reference line over the entire prediction range. Only a few high-capacity samples showed relatively larger deviations, reflecting the limited number of extreme adsorption cases available in the curated database. These observations are consistent with the prediction uncertainty discussed in Section 3.6 and indicate that the model accurately captures the dominant adsorption behaviour while exhibiting greater uncertainty for rare extreme conditions. Model II displayed noticeably greater dispersion around the reference line, consistent with its lower coefficient of determination reported previously (Table 10). Nevertheless, the majority of predictions remained reasonably close to the experimental values, confirming that the ambient-pressure model successfully captured the principal adsorption trends despite the removal of the dominant pressure variation. Overall, the predicted-versus-experimental comparison demonstrates the high predictive accuracy of Model I and confirms that the lower performance of Model II primarily reflects the limited information content available under ambient-pressure conditions rather than deficiencies in the machine learning algorithms (Rudin et al., 2022).

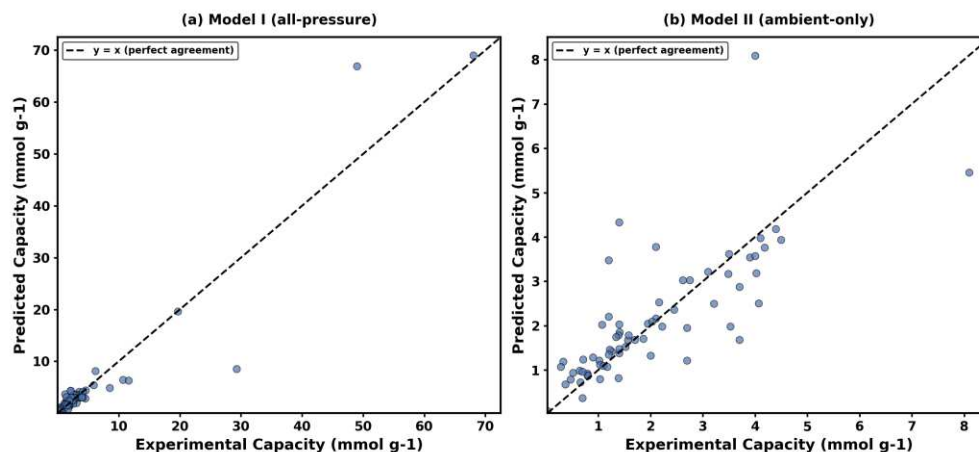


Figure 13. Comparison between experimentally measured and machine-learning-predicted CO₂ adsorption capacities for (a) Model I and (b) Model II. The dashed line represents perfect agreement ($y = x$).

3.7.2 Residual Analysis

Residual analysis was performed to investigate the distribution of prediction errors and to identify any systematic prediction bias in the developed machine learning models (Figure 14). Residuals were calculated as the difference between experimentally measured and predicted CO₂ adsorption capacities. For both models, the residuals were randomly distributed around zero throughout most of the prediction range, indicating the absence of systematic overestimation or underestimation. The mean residuals remained close to zero (Model I: $-0.03 \text{ mmol g}^{-1}$; Model II: $-0.07 \text{ mmol g}^{-1}$), confirming that both models are essentially unbiased. However, increasing residual dispersion was observed at higher adsorption capacities, particularly for Model I. This heteroscedastic behaviour is expected considering the strongly right-skewed distribution of adsorption capacities identified in Section 3.2.1, where relatively few observations occupy the extreme high-capacity region. Consequently, predictions associated with these rare adsorption conditions should be interpreted with greater caution than those corresponding to the densely populated region of the dataset. Overall, the residual analysis confirms that the developed models provide stable and unbiased predictions across the majority of the investigated adsorption range (Freiesleben et al., 2022).

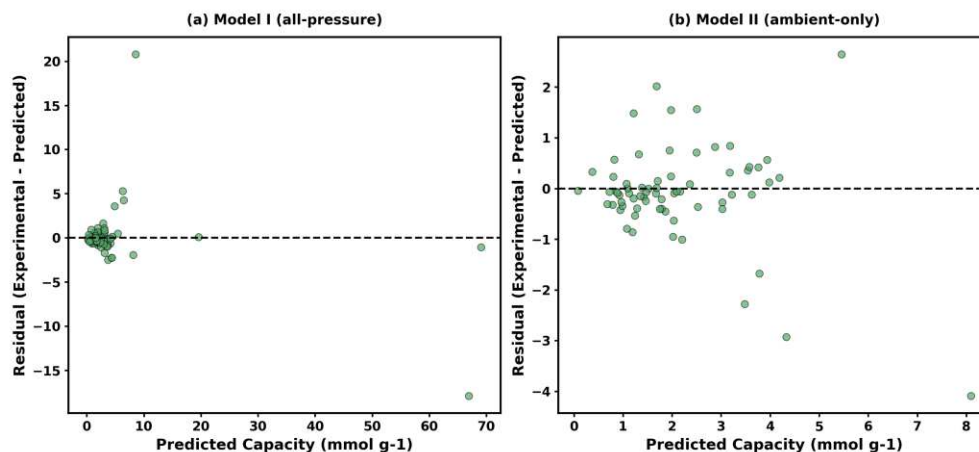


Figure 14. Residual plots (experimental minus predicted capacity versus predicted capacity) for (a) Model I and (b) Model II.

3.7.3 Learning Curve Analysis

Learning curves were constructed to evaluate the relationship between training-set size and predictive performance, thereby distinguishing between limitations arising from insufficient training data and those associated with model complexity (Figure 15). For Model I, the training and cross-validation curves converged progressively with increasing training-set size, reaching final coefficients of determination of approximately 0.994 and 0.954, respectively. The relatively small difference between the two curves indicates excellent generalization with minimal overfitting, suggesting that the current dataset adequately captures the dominant adsorption behaviour represented in the database. In contrast, Model II exhibited a persistent gap between the training and validation curves, with validation performance continuing to improve as additional training samples were incorporated. This behaviour indicates that the ambient-pressure model remains constrained primarily by data availability rather than algorithm selection. The lower predictive performance observed under ambient-pressure conditions therefore reflects the reduced variability of the operating conditions together with the limited availability of complete structural descriptors, rather than an inability of the machine learning algorithms to model the adsorption process. These findings suggest that future expansion of the database through additional well-characterized ambient-pressure adsorption measurements would likely provide greater improvements in predictive performance than further optimization of the regression algorithms. Collectively, the regression diagnostics confirm that the developed models are statistically reliable and free from significant systematic bias. The remaining prediction uncertainty is primarily associated with the limited representation of extreme adsorption conditions rather than deficiencies in the modelling methodology (Geman et al., 1992; Perlich et al., 2003).

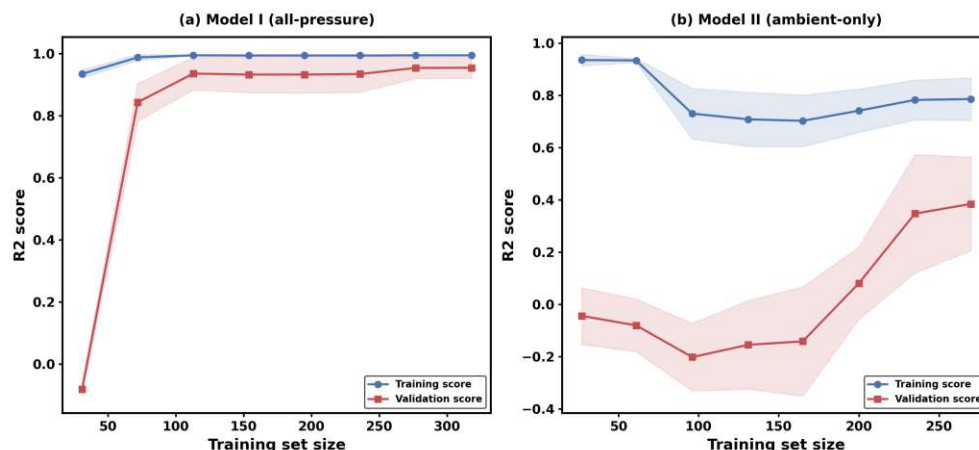


Figure 15. Learning curves (training and cross-validation R^2 versus training-set size) for (a) Model I and (b) Model II (Random Forest).

4. Conclusions

This study developed an explainable machine learning framework to distinguish the effects of operating conditions from intrinsic material properties in CO₂ adsorption, using a curated database of 398 measurements spanning four adsorbent families. Two models were compared: Model I, incorporating the complete pressure range ($R^2 \approx 0.93$), and Model II, restricted to ambient-pressure conditions ($R^2 \approx 0.34$ – 0.42) to reduce pressure confounding. The large gap between these models, together with global and family-specific SHAP analyses, demonstrates that CO₂ partial pressure dominates apparent adsorption performance across the compiled literature, and that BET surface area, pore volume, temperature, and adsorbent family become far more informative once this effect is controlled. Partial dependence analysis further recovered the expected transition from physisorption- to chemisorption-dominated behaviour with increasing temperature, supporting the physical validity of the models. The principal contribution of this work is the quantitative identification of pressure–family confounding in literature-derived CO₂ adsorption databases: family-level performance comparisons that ignore operating conditions can substantially misrepresent intrinsic material behaviour. Because Model II’s predictive accuracy remains modest, its present value lies in identifying which descriptors matter under near-ambient conditions rather than in providing quantitative capacity predictions; expanding well-characterized ambient-pressure data is a priority for future work. More broadly, this study offers a transparent, reproducible framework, combining rigorous data curation, statistical analysis, and explainable AI, for developing physically meaningful predictive models from heterogeneous materials databases.

Data Availability

The curated dataset, the corresponding list of literature sources, and the machine learning scripts developed during this study are available from the corresponding author upon reasonable request.

References:

- Abdelbaky, I., Tayara, H. & Chong, K. T.: 2021, 'Prediction of kinase inhibitors binding modes with machine learning and reduced descriptor sets', *Scientific Reports* 11, 706.
- Abdi, J., Hadavimoghaddam, F., Hadipoor, M. & Hemmati-Sarapardeh, A.: 'Modeling of CO₂ adsorption capacity by porous metal organic frameworks using advanced decision tree-based models'.
- Abdi, J., Hadavimoghaddam, F., Hadipoor, M. & Hemmati-Sarapardeh, A.: 2021, 'Modeling of CO₂ adsorption capacity by porous metal organic frameworks using advanced decision tree-based models', *Scientific reports* 11, 24468.
- Al-Mamoori, A., Saed, U. A., Saoud, A. A., Abdul Jabbar, M. F., Jasim, A., Majdi, H. S., Hanif, A. & Iacomi, P.: 2025, 'Modification of the zeolite ZSM-5 adsorbent for CO₂ capture', *ACS omega* 10, 20129-20141.
- Al-Sakkari, E. G., Ragab, A., So, T. M. Y., Shokrollahi, M., Dagdougui, H., Navarri, P., Elkamel, A. & Amazouz, M.: 'Machine learning-assisted selection of adsorption-based carbon dioxide capture materials'.
- Alatefi, S., Agwu, O. E., Amar, M. N. & Alkouh, A.: 'Comparative analysis of robust machine learning models for predicting CO₂ uptake in porous liquids: Insights for environmental applications'.
- Barnes, B. C., Elton, D. C., Boukouvalas, Z., Taylor, D. E., Mattson, W. D., Fuge, M. D. & Chung, P. W.: 2018, 'Machine learning of energetic material properties', *arXiv preprint arXiv:1807.06156*.
- Breiman, L.: 2001, 'Random forests', *Machine learning* 45, 5-32.
- Chen, T. & Guestrin, C.: 'XGBoost'.
- Chicco, D., Warrens, M. J. & Jurman, G.: 2021, 'The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation', *PeerJ computer science* 7, e623.
- Chowdhury, S., Lin, Y., Liaw, B. & Kerby, L.: 'Evaluation of tree based regression over multiple linear regression for non-normally distributed data in battery performance', *IEEE*, pp. 17-25.
- Deng, F., Zhang, S., Cao, H., Zou, Q., Zhu, F., Su, R., Zeng, Z., Li, L., Ma, X. & Wu, Y.: 'Unraveling CO₂ capture mechanisms in porous carbons across different pressure ranges: A synergistic approach combining machine learning and molecular simulation'.
- Fisher, A., Rudin, C. & Dominici, F.: 2019, 'All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously', *Journal of Machine Learning Research* 20, 1-81.
- Freiesleben, T., König, G., Molnar, C. & Tejero-Cantero, A.: 2022, 'Scientific inference with interpretable machine learning: Analyzing models to learn about real-world phenomena', *arXiv preprint arXiv:2206.05487*.
- Friedman, J. H.: 'Greedy function approximation: A gradient boosting machine'.

Gautam, Sah, R. P. & Sahoo, S.: 2023, 'A review on adsorption isotherms and kinetics of CO₂ and various adsorbent pairs suitable for carbon capture and green refrigeration applications', *Sādhanā* 48, 27.

Geman, S., Bienenstock, E. & Doursat, R.: 1992, 'Neural networks and the bias/variance dilemma', *Neural computation* 4, 1-58.

Geng, N., Liu, N., Chu, S., Huang, Y., Bai, L., Yao, M.-S., Guo, Y. & Zhu, T.: 'Flexible Metal–Organic Frameworks for Gas Handling Operations of CO₂ and Its Isotopes: Mechanisms, Regulation Strategies and Potential Applications'.

Geurts, P., Ernst, D. & Wehenkel, L.: 'Extremely randomized trees'.

Ghaedi, M.: 2021, *Adsorption: Fundamental processes and applications*, Academic press.

Ghawche, R., Grekov, D., Lacroix, C., Poncet, S. & Pré, P.: 2026, 'Screening of adsorbent materials for small-scale compressed CO₂ energy storage using pressure swing adsorption', *Journal of Energy Storage* 143, 119732.

Gunathilake, C. & Jaroniec, M.: 2016, 'Mesoporous calcium oxide–silica and magnesium oxide–silica composites for CO₂ capture at ambient and elevated temperatures', *Journal of Materials Chemistry A* 4, 10914-10924.

Guo, Z., Huang, L., Yan, J., Liu, Y., Guo, Y., Wang, Z., Li, Q., Li, M., Xing, X., Li, L. & Zhang, H.: 'Machine Learning Simulating and Predicting the Adsorption Performance of Activated Carbons for Removing Methylene Blue from Wastewater'.

Hajiali, F., Ellis, N. & Gopaluni, B.: 2025, 'From biomass waste to CO₂ capture: a multi-fidelity machine learning workflow for high-throughput screening of activated carbons', *npj Computational Materials*.

Ibrahim, A. F. & Hussein, M. A.: 'Machine learning predictive models of CO₂ adsorption in sustainable waste-derived activated carbon'.

Iddrisu, M., Umar, A. A., Onaizi, S., Abdulraheem, A., Quddus, M. R. & Hossain, M. M.: 2025, 'Machine learning for data-driven insights into CO₂ adsorption on amorphous porous organic polymers', *Next Materials* 9, 101119.

Iyola, Z., Brantson, E. T., Okeke, N. J., Sanni, K. & Longe, P.: 2025, 'Carbon capture using metal organic frameworks (MOFs): novel custom ensemble learning models for prediction of CO₂ adsorption', *Processes* 13, 2199.

James, G.: 2013, 'An introduction to statistical learning with applications in R', Springer.

Jia, X., Chen, L., Xiong, K., Wang, Y., Zhou, L., Xu, X., Li, S., Wang, M., Thomas, A. & Cheng, C.: 'Machine learning-accelerated discovery of covalent organic frameworks for hydrogen peroxide photosynthesis'.

Jolliffe, I. T. & Cadima, J.: 2016, 'Principal component analysis: a review and recent developments', *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* 374, 20150202.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T.-Y.: 2017, 'Lightgbm: A highly efficient gradient boosting decision tree', *Advances in neural information processing systems* 30.

Kou, J., Liu, T., Jiang, G., Zhou, Z., Yang, G., Pan, Y., Ni, X. & Liu, H.: 2025, 'Interpretable Machine Learning Assisted Screening of Metal-Organic Frameworks for Methane Adsorption Mechanism Exploration under Low and High Pressures', *Journal of Environmental Chemical Engineering*, 120263.

Kruskal, W. H. & Wallis, W. A.: 1952, 'Use of ranks in one-criterion variance analysis', *Journal of the American statistical Association* 47, 583-621.

Kumar, S., Prasad, K., Gil, J. M., Sobral, A. J. F. N. & Koh, J.: 2018, 'Mesoporous zeolite-chitosan composite for enhanced capture and catalytic activity in chemical fixation of CO₂', *Carbohydrate polymers* 198, 401-406.

Levene, H.: 1960, 'Robust tests for equality of variances', *Contributions to probability and statistics*, 278-292.

Li, X., Zhang, X., Zhang, J., Gu, J., Zhang, S., Li, G., Shao, J., He, Y., Yang, H. & Zhang, S.: 2023, 'Applied machine learning to analyze and predict CO₂ adsorption behavior of metal-organic frameworks', *Carbon Capture Science & Technology* 9, 100146.

Little, R. J. A. & Rubin, D. B.: 2019, *Statistical analysis with missing data*, John Wiley & Sons.

Longe, P. O., Davoodi, S., Mehrad, M. & Wood, D. A.: 2025, 'Robust machine-learning model for prediction of carbon dioxide adsorption on metal-organic frameworks', *Journal of Alloys and Compounds* 1010, 177890.

Lundberg, S. M. & Lee, S.-I.: 2017, 'A unified approach to interpreting model predictions', *Advances in neural information processing systems* 30.

Mehrmohammadi, P. & Ghaemi, A.: 2023, 'Investigating the effect of textural properties on CO₂ adsorption in porous carbons via deep neural networks using various training algorithms', *Scientific Reports* 13, 21264.

Nait Amar, M., Alatefi, S., Alkouh, A. & Agwu, O. E.: 'Modeling CO₂ adsorption capacity of diverse porous adsorbent materials: robust machine learning frameworks and insights from local and global explainable artificial intelligence'.

Namdeo, S., Srivastava, V. C. & Mohanty, P.: 2023, 'Machine learning implemented exploration of the adsorption mechanism of carbon dioxide onto porous carbons', *Journal of Colloid and Interface Science* 647, 174-187.

Nunes, L. J. R.: 'The Rising Threat of Atmospheric CO₂: A Review on the Causes, Impacts, and Mitigation Strategies'.

Ojong, O. E., Adoga, O. D., Anaba, U. C., Samuel, S. S., Ana, R. R. & Osha, O. A.: 2026, 'Kinetic Studies and Statistical Design of CO₂ Capture Process from Stream of Gas Mixture at Elevated Temperature Using Composites of Nanoparticles', *Advanced Materials Research* 1188, 131-146.

Oviedo, F., Ferres, J. L., Buonassisi, T. & Butler, K. T.: 2022, 'Interpretable and explainable machine learning for materials science and chemistry', *Accounts of Materials Research* 3, 597-607.

O'Brien, R. M.: 2007, 'A caution regarding rules of thumb for variance inflation factors', *Quality & quantity* 41, 673-690.

Perera, M. S. A., Ranjith, P. G., Choi, S. K., Airey, D. & Weniger, P.: 2012, 'Estimation of gas adsorption capacity in coal: a review and an analytical study', *International Journal of Coal Preparation and Utilization* 32, 25-55.

Perlich, C., Provost, F. & Simonoff, J. S.: 2003, 'Tree induction vs. logistic regression: A learning-curve analysis', *Journal of Machine Learning Research* 4, 211-255.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A.: 2018, 'CatBoost: unbiased boosting with categorical features', *Advances in neural information processing systems* 31.

Rudin, C.: 2019, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature machine intelligence* 1, 206-215.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L. & Zhong, C.: 2022, 'Interpretable machine learning: Fundamental principles and 10 grand challenges', *Statistic Surveys* 16, 1-85.

Sai Bhargava Reddy, M., Ponnammma, D., Sadasivuni, K. K., Kumar, B. & Abdullah, A. M.: 2021, 'Carbon dioxide adsorption based on porous materials', *RSC advances* 11, 12658-12681.

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K. & Menegaz, G.: 'A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME'.

Sekizkardes, A. K., Wang, P., Hoffman, J., Budhathoki, S. & Hopkinson, D.: 2022, 'Amine-functionalized porous organic polymers for carbon dioxide capture', *Materials Advances* 3, 6668-6686.

Shapiro, S. S. & Wilk, M. B.: 1965, 'An analysis of variance test for normality (complete samples)', *Biometrika* 52, 591-611.

Shi, T., Wu, Z., Wu, Z., Zhang, Q. F., Strømme, M. & Xu, C.: 2024, 'Postsynthetic amine modification of porous organic polymers for CO₂ capture and separation', *Journal of Polymer Science* 62, 1554-1568.

Sovacool, B. K., Del Rio, D. F., Herman, K., Iskandarova, M., Uratani, J. M. & Griffiths, S.: 2024, 'Reconfiguring European industry for net-zero: a qualitative review of hydrogen and carbon capture utilization and storage benefits and implementation challenges', *Energy & Environmental Science* 17, 3523-3569.

Sung, I. T., Cheng, Y.-H., Hsieh, C.-M. & Lin, L.-C.: 'Machine Learning for Gas Adsorption in Metal–Organic Frameworks: A Review on Predictive Descriptors'.

Söğüt, E. G. & Gülcan, M.: 2023, 'Adsorption: basics, properties, and classification', *Adsorption through advanced nanoscale materials*, Elsevier, pp. 3-21.

Wang, M., Zeng, Q., Chen, D., Zhang, Y., Liu, J., Ma, C. & Jia, P.: 'A machine learning feature descriptor approach: Revealing potential adsorption mechanisms for SF₆ decomposition product gas-sensitive materials'.

Yuan, L., Zhang, H., Yu, H., Xu, R., Zhang, W., Zhang, Y., Hua, M., Lv, L. & Pan, B.: 'Machine Learning Reveals Key Adsorption Mechanisms for Oxyanions Based on Combination of Experimental and Published Literature Data'.

Yuan, X., Suvarna, M., Low, S., Dissanayake, P. D., Lee, K. B., Li, J., Wang, X. & Ok, Y. S.: 2021, 'Applied machine learning for prediction of CO₂ adsorption on biomass waste-derived porous carbons', *Environmental Science & Technology* 55, 11925-11936.

Zentou, H., Aliyu, M., Abdalla, M. A., Abdelaziz, O. Y., Hoque, B., Alloush, A. M., Tayeb, I. M., Patchigolla, K. & Abdelnaby, M. M.: 2025, 'Advancements and challenges in adsorption-based carbon capture technology: from fundamentals to deployment', *The Chemical Record* 25, e202400188.

Zheng, H., He, Y., Zhu, Y., Liu, L. & Cui, X.: 2021, 'Novel procedure of CO₂ capture of the CaO sorbent activator on the reaction of one-part alkali-activated slag', *RSC advances* 11, 12476-12483.

Zhu, X., Tsang, D. C. W., Wang, L., Su, Z., Hou, D., Li, L. & Shang, J.: 'Machine learning exploration of the critical factors for CO₂ adsorption capacity on porous carbon materials at different pressures'.

Zhu, X., Tsang, D. C. W., Wang, L., Su, Z., Hou, D., Li, L. & Shang, J.: 2020, 'Machine learning exploration of the critical factors for CO₂ adsorption capacity on porous carbon materials at different pressures', *Journal of Cleaner Production* 273, 122915.