Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Integrating Molecular Simulations and Machine Learning for Predicting HFC Adsorption in Zeolites

Abrar A. Elhussien
King Fahd University of Petroleum and Minerals
Nourhan Salem
King Fahd University of Petroleum and Minerals
Mustafa M. Amin
aminmustafa289@gmail.com

King Fahd University of Petroleum and Minerals

Research Article

Keywords: Adsorption, Hydrofluorocarbons, Zeolites, Molecular simulation, Machine learning, Ensemble models, SHAP, Model validation

Posted Date: August 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10707831/v1

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

Additional Declarations: No competing interests reported.

Integrating Molecular Simulations and Machine Learning for Predicting HFC Adsorption in Zeolites

Abrar A. Elhussien*, Nourhan Salem, and Mustafa M. Amin*

Chemical Engineering Department, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia

Abstract

This study re-audits and models a simulation-derived database for the adsorption of fluoromethane (R41), difluoromethane (R32), trifluoromethane (R23), and tetrafluoromethane (R14) in zeolites 13X, 3A, 4A, and 5A. The fully documented primary population contains 1,536 adsorption-loading observations (4 HFCs × 4 zeolites × 6 temperatures × 16 pressures) and 1,440 modeled isosteric-heat values; the 96 source heat zeros occurring at P = 0 were preserved in the audit data but were not used as evaluated heat targets. Ensemble regressors were compared under leakage-aware validation. Model selection was performed only within the development pool using five-fold isotherm-grouped cross-validation, followed by evaluation on a predefined balanced HFC–zeolite-pair holdout. Random row-level splitting was retained only as an interpolation benchmark. Histogram Gradient Boosting was selected for adsorption loading and achieved $R^2 = 0.845$, RMSE = 0.512 mmol g^{-1}, and MAE = 0.371 mmol g^{-1} on the balanced pair holdout. Extra Trees was selected for isosteric heat and achieved $R^2 = 0.885$, RMSE = 0.617 kJ mol^{-1}, and MAE = 0.559 kJ mol^{-1}. Model attribution indicates that pressure and temperature are important for loading, whereas heat prediction depends more strongly on correlated framework and molecular descriptor families. Leave-one-HFC-out and leave-one-zeolite-out analyses show that the present surrogate is reliable mainly for interpolation within the studied chemical domain and is not yet validated for unseen HFC or zeolite identities. The revised workflow therefore provides a transparent and reproducible surrogate while defining its applicability limits.

Keywords: Adsorption; Hydrofluorocarbons; Zeolites; Molecular simulation; Machine learning; Ensemble models; SHAP; Model validation.

1. Introduction

Hydrofluorocarbons (HFCs), such as fluoromethane (R41), difluoromethane (R32), trifluoromethane (R23), and tetrafluoromethane (R14), are widely utilized in refrigeration, air conditioning, and foam-blowing industries due to their excellent thermodynamic properties and safety profile (Ravishankara et al. 2009; Myhre et al. 2014). However, HFCs are potent greenhouse gases (GHGs) with high global warming

potentials (GWPs), necessitating urgent efforts to reduce their environmental footprint (Purohit and Höglund-Isaksson 2017). International agreements, such as the Kigali Amendment to the Montreal Protocol, aim to phase down HFC emissions and promote alternative solutions (Baker and Lokhandwala 2008). This necessitates the development of efficient capture and separation techniques for HFCs to mitigate their environmental impact.

Conventional separation techniques, including cryogenic distillation and chemical absorption, face significant limitations. Cryogenic distillation, while effective, is energy-intensive and costly, particularly for separating HFCs with similar boiling points or azeotropic behavior (Castle 2002; De Monte 2002). Similarly, chemical absorption methods involve complex operational challenges and high maintenance costs (Fang et al. 2022; Wang et al. 2011). Adsorption-based processes using porous materials, such as zeolites, have emerged as a viable alternative due to their energy efficiency, reusability, and selective adsorption capabilities (Taddei and Petit 2021; Li et al. 2009; Pérez-Botella et al. 2022; Ma et al. 2020; Cejka et al. 2010; Khaleque et al. 2020).

Zeolites are crystalline microporous aluminosilicates known for their high surface area, thermal stability, and tunable pore structures (Thommes and Cychosz 2014; Yancey et al. 2022b; Ma et al. 2020). Among these, 13X (FAU framework) and 3A, 4A, and 5A (LTA framework) zeolites exhibit promising performance in gas separation due to their unique structural properties. Zeolite 13X, characterized by its large pore size and high void fraction, facilitates strong cation-fluorine interactions, enhancing adsorption selectivity for polar HFC molecules (Wanigarathna 2018; Freni et al. 2015; Elhussien et al. 2025). In contrast, the LTA framework in zeolites 3A, 4A, and 5A provides smaller pore sizes, enabling selective adsorption based on molecular size and kinetic diameter (Sadeghi et al. 2024; Yancey et al. 2022a). These properties make zeolites ideal candidates for separating methane derivatives like HFCs, which vary in molecular weight, dipole moment, and kinetic diameter.

Despite their potential, accurately predicting the adsorption behavior of HFCs in zeolites under varying operational conditions remains challenging due to the complexity of molecular interactions (Yan et al. 2021). Molecular simulations, such as Grand Canonical Monte Carlo (GCMC) and Molecular Dynamics (MD), have been instrumental in elucidating adsorption mechanisms and exploring zeolite frameworks (Mollahosseini and Abdelrasoul 2021; Smit and Maesen 2008; Abdelrasoul et al. 2017; Sun et al. 2024). However, these methods are computationally intensive and limited in scalability, particularly when exploring a wide range of material configurations (Evans and Coudert 2024; Salahshoori et al. 2024b; Salahshoori et al. 2024a).

The advent of artificial intelligence (AI) has transformed the field of adsorption science by offering rapid and accurate predictions. Machine learning (ML) models have demonstrated their ability to capture complex, non-linear relationships between input features and adsorption outputs (Barpaga et al. 2021; Butler et al. 2018; Zhong et al. 2021; Yang et al. 2023). These models can integrate molecular, zeolite, and operational parameters, enabling comprehensive predictions of adsorption capacity and isosteric heat (Chakraborty et al. 2024). Artificial neural networks and deep neural networks excel at capturing intricate molecular-material interactions through their multi-layered architectures, making them effective for predictive tasks involving non-linear data relationships (Yulia et al. 2021; Gupta and Raza 2019; Oliveira et al. 2020). Ensemble learning algorithms provide robust predictions while offering insights into feature importance, making them highly interpretable (Sagi and Rokach 2021; Luo et al. 2023). In addition, CatBoost also with its ability to handle categorical and numerical data efficiently, is particularly well-suited for high-dimensional datasets encountered in adsorption studies (Achour and Hosni 2024).

AI applications in adsorption science are gaining momentum. ANNs have been demonstrated to effectively predict methane adsorption in metal-organic frameworks (MOFs), highlighting their adaptability and accuracy (Li et al. 2024). Similarly, XGBoost and other ML models have been applied to predict gas adsorption in nanoporous materials, achieving significant improvements in computational efficiency (Fanourgakis et al. 2020). Recent advancements also include integrating feature selection methods, such as Shapley Additive Explanations (SHAP), to enhance model interpretability and optimize adsorption processes (Al-Sakkari et al. 2023; Wang et al. 2024).

This study combines the previously generated molecular-simulation adsorption database with a revised machine-learning analysis for R41, R32, R23, and R14 in zeolites 13X, 3A, 4A, and 5A. The present contribution is the data audit, transparent reconstruction of the fully documented 1,536-row population, leakage-aware model development, and applicability analysis. Bagging and boosting ensemble methods are compared, but final model selection is based on isotherm-grouped cross-validation within a development pool rather than on a row-random test split. A balanced HFC–zeolite-pair holdout is then used to test transfer to new combinations of HFC and zeolite identities already represented during development. Feature attribution is interpreted as model attribution rather than causal physical proof because molecular and zeolite descriptors are strongly identity-locked in this four-by-four chemical space.

The paper is organized as follows: Section 2 illustrates the data generation and analysis. While Section 3 discusses the ML modeling framework and the used ML algorithms, Section 4 presents the findings and obtained results. Section 5 concludes the paper and proposes adirection for future work.

2. Data Generation

2.1 Adsorption Data from Molecular Simulations

The adsorption targets analyzed in the present work were inherited from the previously published molecular-simulation study by Elhussien et al. (2025); they were not newly generated by the present ML analysis. In that study, Grand Canonical Monte Carlo (GCMC) and Molecular Dynamics (MD) simulations were used to evaluate adsorption loading and adsorption energetics for HFCs in zeolite structures. The published simulation protocol employed the Consistent Valence Force Field (CVFF), Ewald summation for electrostatic interactions, and a Lennard-Jones treatment for van der Waals interactions. The prior work compared the molecular-simulation protocol with available experimental adsorption data. In the present study, the raw numerical adsorption targets are preserved, while the new contribution is the audit, curation, machine-learning analysis, validation strategy, and model interpretation.

The selected zeolite structures, 3A, 4A, 5A, and 13X, represent different pore/window environments and cation compositions. Crystallographic Information Files were processed in Materials Studio and represented as 2×2×2 supercells in the source simulation work. The structural models shown in Figure 1 are retained from the original manuscript because they directly illustrate the LTA and FAU frameworks used to generate the simulation database. The source-reported structural descriptors used in the manuscript include PLD, LCD, zeolite/framework density, void fraction (VF), volumetric surface area (VSA), gravimetric surface area (GSA), and specific pore volume (Vp), summarized in Table 1. For the revised ML audit, these source values are treated as material descriptors rather than as independently varying causal variables, because only four zeolite identities are represented.

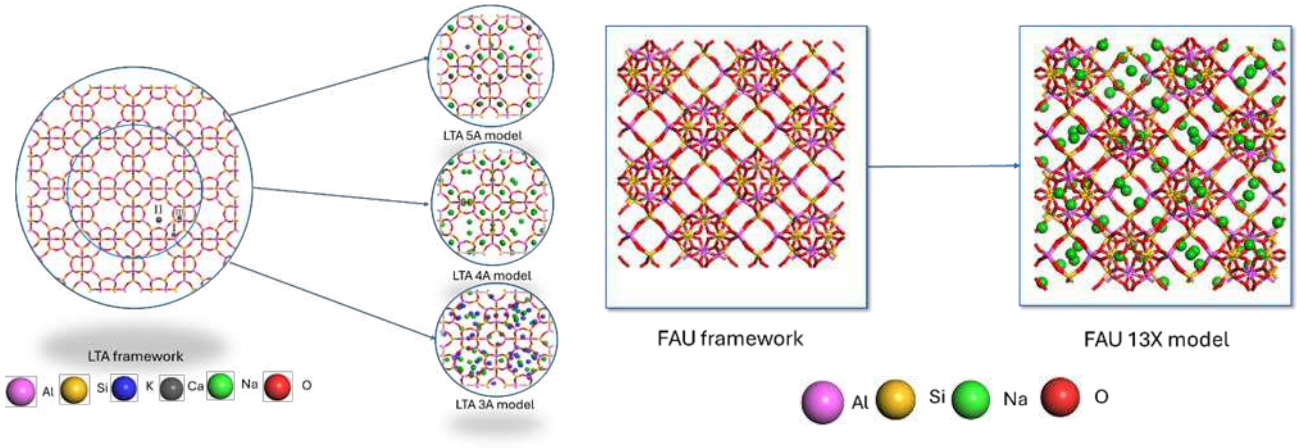


Figure 1. Structures of the LTA and FAU zeolite frameworks used in the simulation database.

Table 1. Zeolite frameworks and source-reported structural properties.

Zeolite	PLD (Å)	LCD (Å)	DZ (g/cm³)	VF	VSA (m²/cm³)	GSA (m²/g)	Vp (cm³/g)
13X (Na88Al88Si104O384)	10.6957	10	1.425	0.2700	2263.11	3225.45	0.389
5A (Na32Ca32Al96Si96O384)	10.22552	5	1.498	0.4388	2822.43	4227.99	0.658
4A (Na96Al96Si96O384)	10.22552	4	1.519	0.3468	2694.63	4094.16	0.527
3A (K58Na38Al96Si96O384)	10.22552	3	1.623	0.1273	1748.33	2837.74	0.306

2.2 Properties of Hydrofluorocarbons (HFCs)

The study includes the methane-derived HFC refrigerants R41, R32, R23, and R14. Molecular descriptors are used to distinguish the four adsorbates and to examine adsorption trends. Dipole moment, normal boiling point, and molecular weight were audited against the source manuscript and standard property references; kinetic diameter and molecular density are retained here as descriptive properties because they are used in the original figures. In the leakage-aware ML reconstruction, the chemically transparent stoichiometric descriptors H count and F count were also used, while conclusions based on any single molecular descriptor are kept cautious because all molecular properties are locked to only four HFC identities. The properties retained in the original manuscript format are listed in Table 2.

Table 2. Properties of the selected HFCs.

HFC	D(D)	BP (K)	MW (g/mol)	KD (Å)	DM (g/cm³)
R41	1.896	194.9	34.0329	3.45	0.00144
R32	1.972	221.4	52.0234	3.80	0.00181
R23	1.594	191.0	70.0138	3.58	0.00156
R14	0.000	145.1	88.0043	3.92	0.00196

2.3 Data Descriptive Analysis

The supplied headerless workbook contains 3,072 rows. Audit of the row structure identified a fully documented primary block of 1,536 observations corresponding to 4 HFCs × 4 zeolites × 6 temperatures × 16 pressures. The remaining 1,536 rows are preserved separately for auditability but are excluded from the

primary inferential analysis because explicit HFC identities and complete descriptor provenance are not present in the supplied source files. All 1,536 loading values in the primary block are retained. For isosteric heat, the source file contains 96 zero values, all at P = 0; these entries are preserved in the raw heat column but treated as unavailable in the modeled heat target, leaving 1,440 heat observations. Table 3 summarizes the audited primary population.

The audited loading values range from 0 to 5.22 mmol g^{-1}, with a mean of 2.381 mmol g^{-1} and a standard deviation of 1.157 mmol g^{-1}. Valid modeled isosteric heats range from 5.34 to 12.98 kJ mol^{-1}, with a mean of 9.466 kJ mol^{-1} and a standard deviation of 1.792 kJ mol^{-1}. The operating space spans 293–333 K and 0–100 kPa. Because the four HFCs and four zeolites contribute balanced numbers of operating points, global model fitting is not dominated by one HFC–zeolite pair; however, repeated pressure–temperature conditions within the same pair are highly related and therefore require grouped rather than row-random validation for an honest generalization estimate.

Table 3. Descriptive statistics of the audited primary dataset.

Feature	Max	Min	Mean	Std. Dev.	Skewness	Kurtosis
D (D)	1.972	0.0000	1.365	0.8012	-1.053	-0.7546
BP (K)	221.40	145.10	188.10	27.452	-0.5114	-0.9494
MW (g/mol)	88.004	34.033	61.019	20.120	0.0000	-1.361
KD (Å)	3.920	3.450	3.688	0.1836	-0.0262	-1.590
Molecular density (g/cm³)	0.00196	0.00144	0.00169	0.00020	0.0688	-1.607
PLD/source descriptor (Å)	10.696	10.226	10.343	0.2037	1.156	-0.6649
LCD/window descriptor (Å)	10	3	5.500	2.693	0.9230	-0.8383
Framework density (g/cm³)	1.623	1.425	1.516	0.0708	0.3175	-1.019
VF	0.4388	0.1273	0.2957	0.1142	-0.2916	-1.187
VSA (m²/cm³)	2822.43	1748.33	2382.12	420.67	-0.4724	-1.332
GSA (m²/g)	4227.99	2837.74	3596.34	583.25	-0.1416	-1.767
Vp (cm³/g)	0.6580	0.3060	0.4700	0.1343	0.1954	-1.440
T (K)	333	293	310.50	14.073	0.3370	-1.289
P (kPa)	100.00	0.0000	50.000	30.742	0.0000	-1.210
Loading (mmol/g)	5.220	0.0000	2.381	1.157	0.0383	-0.0920
Isosteric heat (kJ/mol), N=1,440	12.980	5.340	9.466	1.792	0.3251	-0.8452

Figure 2 is retained from the original manuscript as an exploratory correlation visualization of molecular, zeolite, and operating descriptors. It is useful for showing the strong collinearity within molecular-property and zeolite-property groups, but simple pairwise correlations should not be interpreted causally because only four HFC identities and four zeolite identities generate most descriptor variation. In addition, the

revised heat model excludes the 96 P = 0 source-zero entries; therefore, the correlation map is treated as descriptive context rather than as the basis for feature selection or mechanistic attribution. Pressure shows the expected positive association with loading, while temperature generally acts in the opposite direction for adsorption loading over the sampled range.

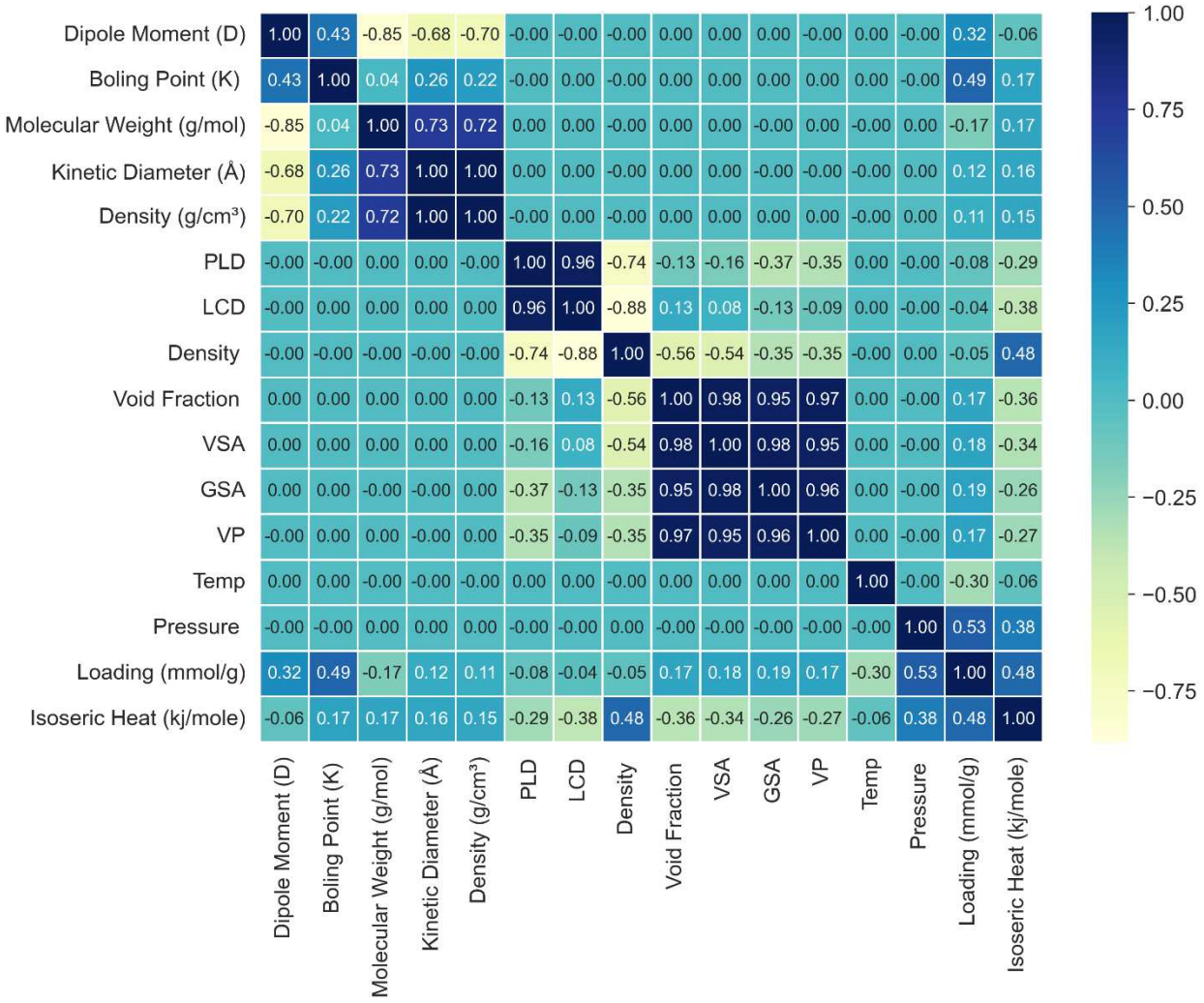


Figure 2. Correlation analysis of molecular, zeolite, and operating descriptors in the original exploratory representation. The map is descriptive; correlated identity-locked descriptors are not interpreted as causal effects.

3. ML Modeling Framework

3.1 Approach Overview

The revised machine-learning framework predicts two targets: adsorption loading and isosteric heat. Input variables describe the HFC identity, zeolite identity, and operating state. To preserve the original manuscript structure, the descriptor groups are summarized below, while the revised modeling code uses

an audited set of molecular, zeolite, temperature, and pressure descriptors and records the HFC–zeolite pair and isotherm identities only for grouping and validation.

- Zeolite properties: source-reported PLD/cavity descriptor, pore/window opening descriptor, framework density, VF, VSA, GSA, and Vp.
- HFC properties: dipole moment, boiling point, molecular weight, and stoichiometric H/F counts; kinetic diameter is retained for descriptive analysis in the original figures.
- Operational conditions: temperature and pressure.

The analysis began with source-integrity checks, descriptive statistics, and correlation analysis. Each observation was assigned an HFC–zeolite pair identifier and an isotherm identifier (HFC–zeolite–temperature). These grouping variables were not used as predictive features; they were used to prevent closely related points from the same isotherm from being distributed across development folds.

The primary loading dataset contains 1,536 observations. A predefined balanced pair holdout contains four HFC–zeolite combinations—R41–13X, R32–5A, R23–4A, and R14–3A—giving 384 loading observations and 360 valid heat observations for final evaluation. The remaining 12 HFC–zeolite pairs form the development pool (1,152 loading and 1,080 heat observations). Model-family and hyperparameter selection were performed only in this development pool using five-fold isotherm-grouped cross-validation. The pair holdout was not used during model selection. This design tests transfer to new combinations of HFC and zeolite identities that were individually seen during development, but it does not constitute validation on an entirely unseen HFC chemistry or unseen zeolite framework.

For comparison with the original analysis, row-level 80/20 splits were also repeated as an interpolation benchmark. Such splits place neighboring operating conditions from the same HFC–zeolite system in both training and test sets and therefore produce much higher apparent accuracy. Figure 3 is intentionally retained from the original manuscript to illustrate this conventional row-random workflow; in the revised study it is treated as a reference benchmark rather than as the final generalization test.

Additional leave-one-pair-out, leave-one-HFC-out, and leave-one-zeolite-out analyses were used to define the applicability domain. These tests are reported to distinguish interpolation among known chemical identities from extrapolation to an unseen HFC or zeolite identity.

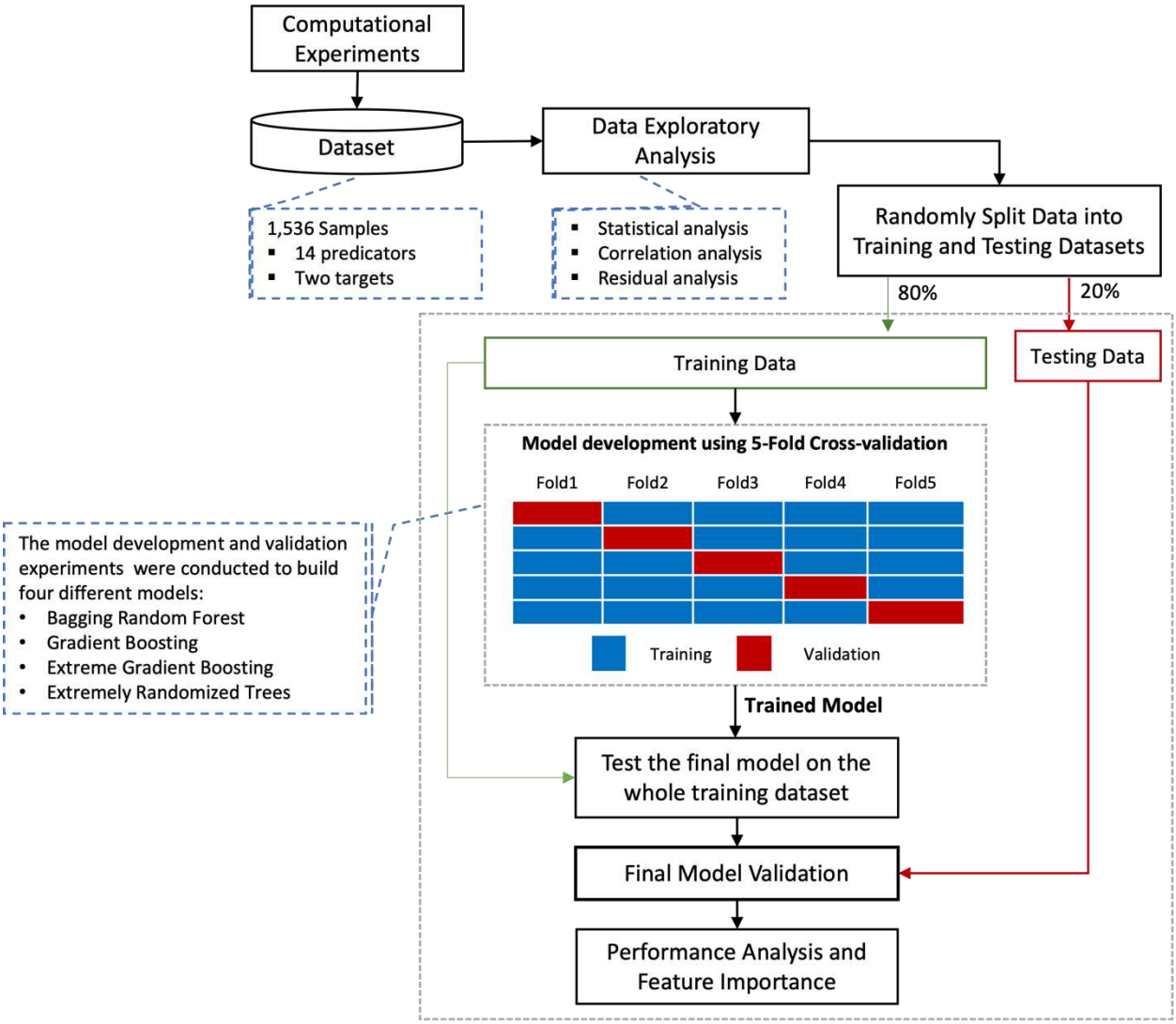


Figure 3. Original row-random ML workflow retained as an interpolation benchmark. Final reported generalization uses development-only isotherm-grouped cross-validation followed by a predefined balanced HFC–zeolite-pair holdout.

3.2 Ensemble Learning Methods

Ensemble learning methods improve non-linear prediction by aggregating multiple decision trees or sequentially correcting residual errors. The original analysis considered Random Forest, Extra Trees Regressor, Gradient Boosting, XGBoost, and CatBoost. For the revised fully reproducible comparison, Ridge regression was used as a linear baseline and Random Forest, Extra Trees, Gradient Boosting, and Histogram Gradient Boosting were evaluated under the same leakage-aware validation protocol. The original Random Forest, Gradient Boosting, CatBoost, and Extra Trees parity figures are retained later as the row-random interpolation benchmark, while final model selection is based on grouped development performance.

3.2.1 Bagging-Based Models

Random Forest (RF) and Extra Trees Regressor (ETR) construct ensembles of decision trees and average their predictions, thereby reducing variance relative to a single tree. RF introduces randomness through bootstrap sampling and random feature subsets, whereas ETR adds additional randomization in the split thresholds. Both algorithms are well suited to non-linear descriptor interactions but can nearly memorize dense simulation grids if assessed only on neighboring row-level test points; this is why grouped validation is essential in the present work.

For tree-based regression, model selection minimizes prediction error on held-out development folds, with RMSE used as the primary tuning criterion.

$$\mathrm{MSE} = (1/n)\,\Sigma_{i=1}^{n}\,(y_i - \hat{y}_i)^2$$

For observations y_i and predictions $\hat{y}_i$, the squared-error contribution is $(y_i - \hat{y}_i)^2$; fold-level RMSE is obtained from the mean squared error across the held-out fold.

The number of trees, maximum depth or leaf complexity, and minimum samples per split/leaf were tuned only within the development pool. Training-set fit is reported as a diagnostic, but near-zero training error for highly flexible tree ensembles is interpreted as a memorization warning rather than as evidence of superior generalization.

3.2.2 Boosting-Based Models

Gradient Boosting builds trees sequentially so that each new learner reduces the residual error of the preceding ensemble. The original analysis also considered XGBoost and CatBoost. In the revised reproducible workflow, conventional Gradient Boosting and Histogram Gradient Boosting were compared. Histogram Gradient Boosting groups continuous features into bins and provides an efficient non-linear learner for the dense simulation grid while retaining regularization through learning rate, leaf size, and iteration controls.

The boosting objective can be written generally as the sum of a prediction loss and a model-complexity or regularization contribution.

$$L = \Sigma_i\, l(y_i, \hat{y}_i) + \Sigma_k\, \Omega(f_k)$$

Here the loss compares the simulation-derived reference target with the model prediction, while regularization and tree-complexity controls limit overfitting. Hyperparameters were selected from the development data only and were fixed before evaluation of the balanced pair holdout.

The final loading model was Histogram Gradient Boosting because it achieved the lowest isotherm-grouped development RMSE among the candidate models. The final heat model was Extra Trees. This target-specific selection was performed before examining the balanced pair holdout results.

3.3 Performance Evaluation

Model performance was quantified using the coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE). Bias and range-normalized RMSE were also recorded in the reproducibility package as secondary diagnostics. Because R^2 can become unstable for very narrow held-out target ranges, RMSE and MAE are emphasized when interpreting leave-one-pair results.

- Mean Absolute Error (MAE):

  $$MAE = (1/n) \Sigma_{i=1}^{n} |y_i - \hat{y}_i|$$

  MAE is the mean absolute difference between the simulation-derived reference values and model predictions.

- Root Mean Squared Error (RMSE):

  $$RMSE = [(1/n) \Sigma_{i=1}^{n} (y_i - \hat{y}_i)^2]^{1/2}$$

  RMSE is the square root of the mean squared prediction error and gives greater weight to larger errors.

- R-Squared (R^2):

  $$R^2 = 1 - [\Sigma_{i=1}^{n} (y_i - \hat{y}_i)^2 / \Sigma_{i=1}^{n} (y_i - \bar{y})^2]$$

  R^2 measures the fraction of target variance explained relative to predicting the mean of the evaluated set. Negative R^2 values are possible when extrapolation performs worse than that mean-reference model.

  $$\bar{y} = (1/n) \Sigma_{i=1}^{n} y_i$$

Five-fold isotherm-grouped cross-validation was used for model development. Final performance was then calculated once on the predefined balanced HFC–zeolite-pair holdout. The original row-level train/test metrics and parity figures are retained only as an interpolation benchmark so that the effect of validation strategy can be seen directly.

4. Results and Discussion

This section first retains the original adsorption-distribution and descriptor figures to describe the simulation database, then distinguishes the high interpolation accuracy obtained by row-random splitting from the more demanding leakage-aware pair-holdout performance. The discussion therefore separates descriptive adsorption trends, predictive performance, and model attribution rather than treating correlated descriptors as independent physical causes.

4.1 Adsorption Behavior of HFCs Across Zeolites

The following subsections examine the adsorption behavior of the four HFCs across 13X, 5A, 4A, and 3A using the original distribution and descriptor figures. Because each panel aggregates multiple pressures and temperatures, the plots describe the sampled database and should not be read as fixed-condition selectivity rankings.

4.1.1 Adsorption Capacity and Frequency of HFCs Across Zeolites

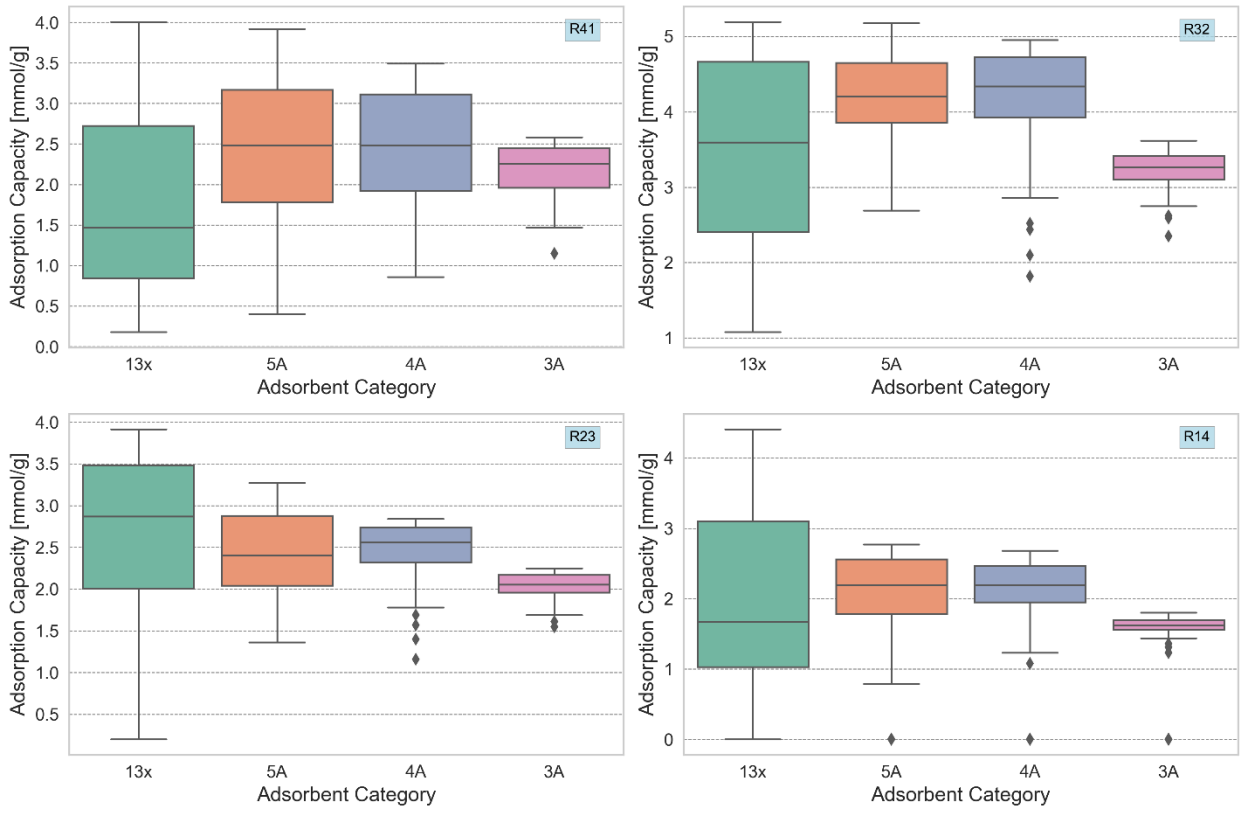


Figure 4. Adsorption-loading distributions for the four HFCs across zeolite types (original figure retained).

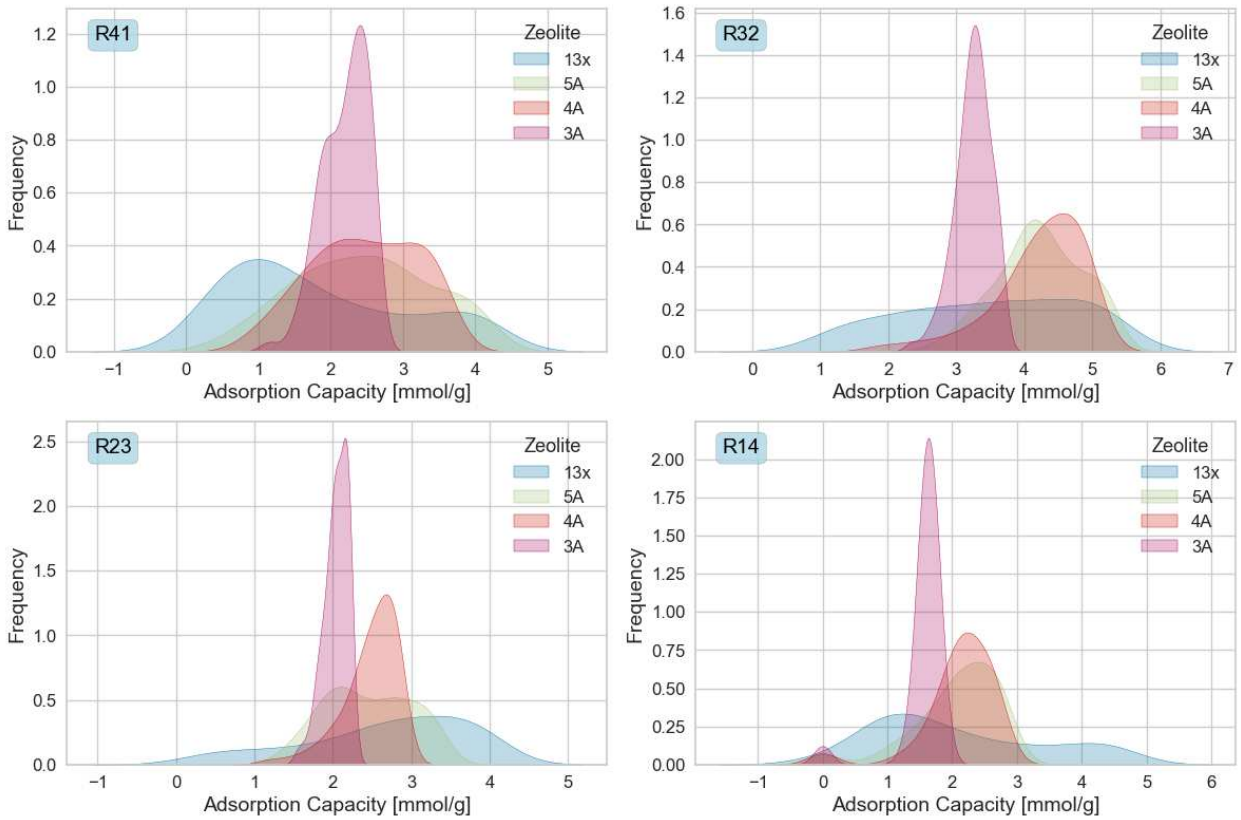


Figure 5. Frequency distributions of adsorption loading for the four HFCs across zeolite types (original figure retained).

Figures 4 and 5 show strongly HFC-dependent loading distributions rather than a universal ranking in which one zeolite is always superior. Median loading from the audited primary population is highest for R32 in 4A (4.25 mmol g^{-1}) and 5A (4.14 mmol g^{-1}), whereas R23 has its highest median in 13X (2.63 mmol g^{-1}). R41 and R14 also show higher median loading in some LTA materials than in 13X. The broader distributions of several 13X and 5A systems indicate a stronger response to the sampled pressure–temperature range, while 3A often gives a narrower loading range. These results show that pore accessibility, cation environment, guest polarity, and confinement act jointly; the figures therefore support a system-specific rather than single-descriptor interpretation.

4.1.2 Zeolite Properties and HFC Adsorption Capacity

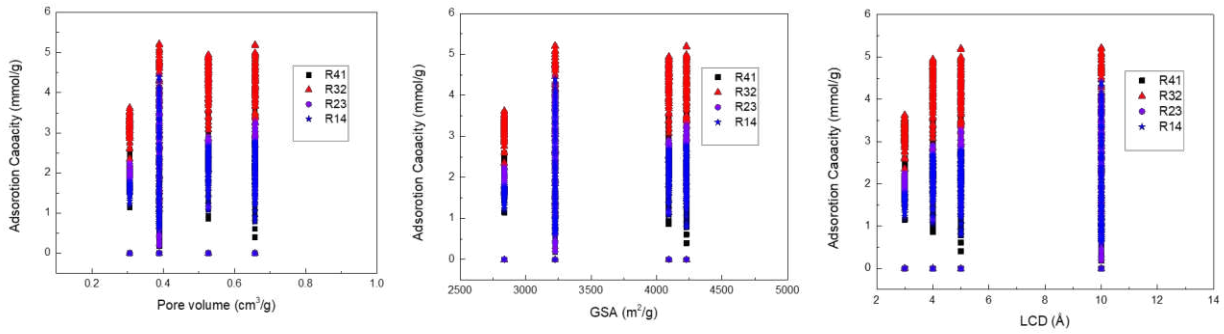


Figure 6. Adsorption loading versus (a) pore volume, (b) GSA, and (c) the source-reported cavity/window descriptor (original figure retained).

Figure 6 demonstrates why simple one-variable structure–property interpretations are insufficient. Only four zeolite identities are represented, so pore volume, GSA, density, void fraction, and pore/window descriptors are strongly correlated with zeolite identity. A larger pore or surface area does not guarantee the highest loading for every HFC. For example, the audited median loading of R32 is higher in 4A and 5A than in 13X, despite the larger FAU cavity of 13X. The observed adsorption therefore reflects coupled effects of accessibility, confinement, cation composition, and HFC identity rather than a monotonic dependence on one geometric descriptor.

The FAU and LTA materials nevertheless provide physically distinct adsorption environments. 13X offers a large supercage and a more open window, whereas the effective LTA opening depends on the exchanged cation. Tight confinement can enhance local host–guest interactions for a molecule that remains accessible, while excessive restriction can limit loading. The structural descriptors in Figure 6 should consequently be viewed as correlated fingerprints of the zeolite systems, not as independent experimental control variables.

4.1.3 HFC Adsorption Capacity and Molecular Properties

Figure 7 retains the original comparison of loading with boiling point, dipole moment, and kinetic diameter. The audited data show that R32, which has the highest listed boiling point and dipole moment, also has the highest overall median loading, whereas nonpolar R14 generally loads less strongly. However, only four HFC identities are present and boiling point, molecular weight, fluorination level, dipole moment, and kinetic diameter co-vary with identity. The plots therefore support the importance of molecular identity and host–guest compatibility, but they do not isolate a causal effect of any single molecular descriptor. This caution is carried forward to the SHAP interpretation below.

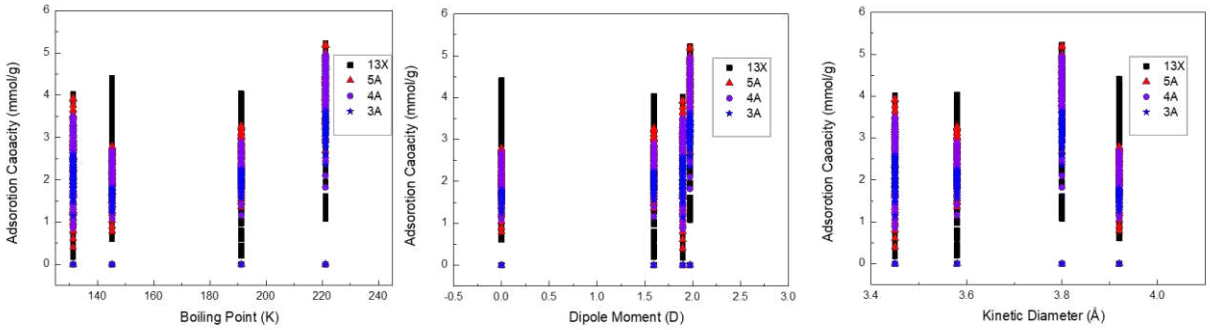


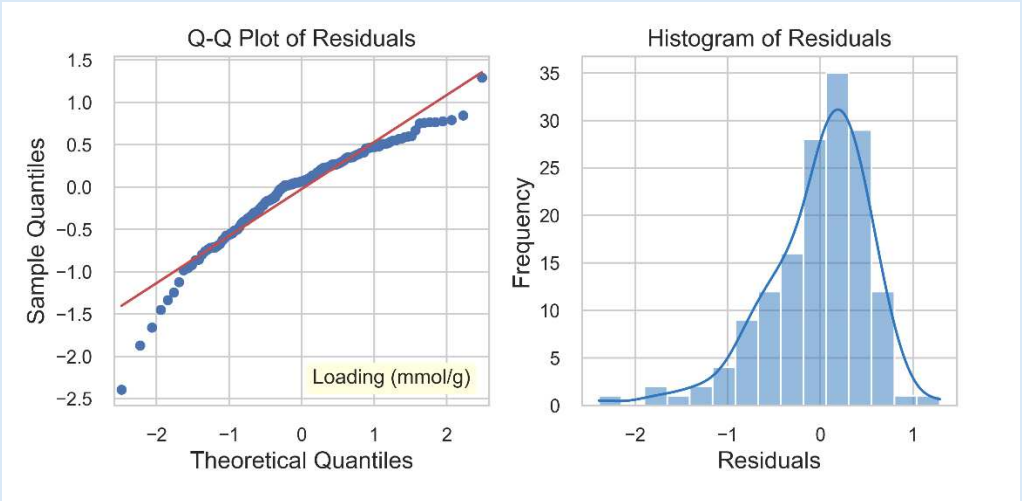
Figure 7. Adsorption loading across zeolites as a function of boiling point, dipole moment, and kinetic diameter (original figure retained; trends are descriptive rather than causal).

4.2 Adsorption Capacity Modeling and Prediction

4.2.1 OLS Regression Baseline Model

Ordinary least-squares (OLS) regression was retained as a transparent linear baseline. Its purpose is not to provide the final adsorption surrogate but to test whether a single linear mapping can reproduce the combined molecular, zeolite, pressure, and temperature effects. The structured simulation grid and correlated descriptor families make substantial non-linearity expected.

The Q–Q plots and residual histograms in Figure 8, retained from the original analysis, show departures from ideal residual normality and pronounced tails, particularly for isosteric heat. These diagnostics support the use of non-linear ensemble models. They do not, by themselves, establish generalization; that question is addressed by the grouped development validation and balanced pair holdout.



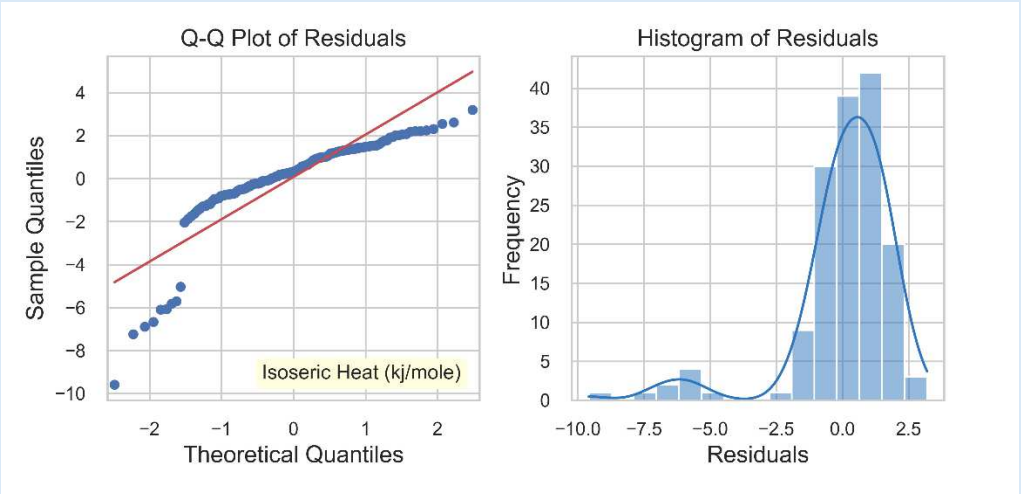


Figure 8. Q–Q plots and residual distributions for the original OLS baseline analysis (original figure retained).

4.2.2 ML Prediction of Isosteric Heat (kJ/mol)

The original row-random benchmark for isosteric heat is retained in Table 4 and Figure 10 to show why conventional random splitting can appear exceptionally accurate. Under that benchmark, the tree ensembles produce R^2 values near 0.97–0.99 because neighboring pressure–temperature points from the same HFC–zeolite systems can occur in both training and testing data. In the revised leakage-aware analysis, Extra Trees was selected using only the development pool. Its five-fold isotherm-grouped development performance was $R^2 = 0.993 \pm 0.006$, RMSE = 0.141 ± 0.054 kJ mol^{-1}, and MAE = 0.082 ± 0.018 kJ mol^{-1}. On the balanced HFC–zeolite-pair holdout, the final heat model achieved $R^2 = 0.885$, RMSE = 0.617 kJ mol^{-1}, and MAE = 0.559 kJ mol^{-1}. These values provide a more conservative estimate of performance when predicting HFC–zeolite combinations withheld from model development.

Table 4. Original row-random interpolation benchmark for isosteric-heat models. These values are retained for comparison and are not the final leakage-aware generalization estimate.

Model	Phase	R2	MSE	MAE	RAE
Bagging: Random Forest	Training	0.998 ± 0.000	0.003 ± 0.000	0.035 ± 0.001	0.038 ± 0.001
	Validation	0.986 ± 0.004	0.019 ± 0.005	0.092 ± 0.010	0.102 ± 0.012
	Testing	**0.989**	**0.014**	**0.080**	**0.095**
Gradient Boosting	Training	0.977 ± 0.001	0.032 ± 0.001	0.135 ± 0.000	0.149 ± 0.003
	Validation	0.967 ± 0.007	0.045 ± 0.007	0.162 ± 0.011	0.179 ± 0.014
	Testing	**0.966**	**0.042**	**0.156**	**0.184**
Advanced: CatBoost	Training	0.994 ± 0.000	0.008 ± 0.000	0.064 ± 0.001	0.070 ± 0.001
	Validation	0.989 ± 0.001	0.015 ± 0.001	0.086 ± 0.003	0.094 ± 0.005
	Testing	**0.990**	**0.012**	**0.078**	**0.092**
Extra Tress Regressor	Training	1.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Validation	0.991 ± 0.002	0.012 ± 0.003	0.072 ± 0.006	0.080 ± 0.008
	Testing	**0.992**	**0.010**	**0.067**	**0.080**

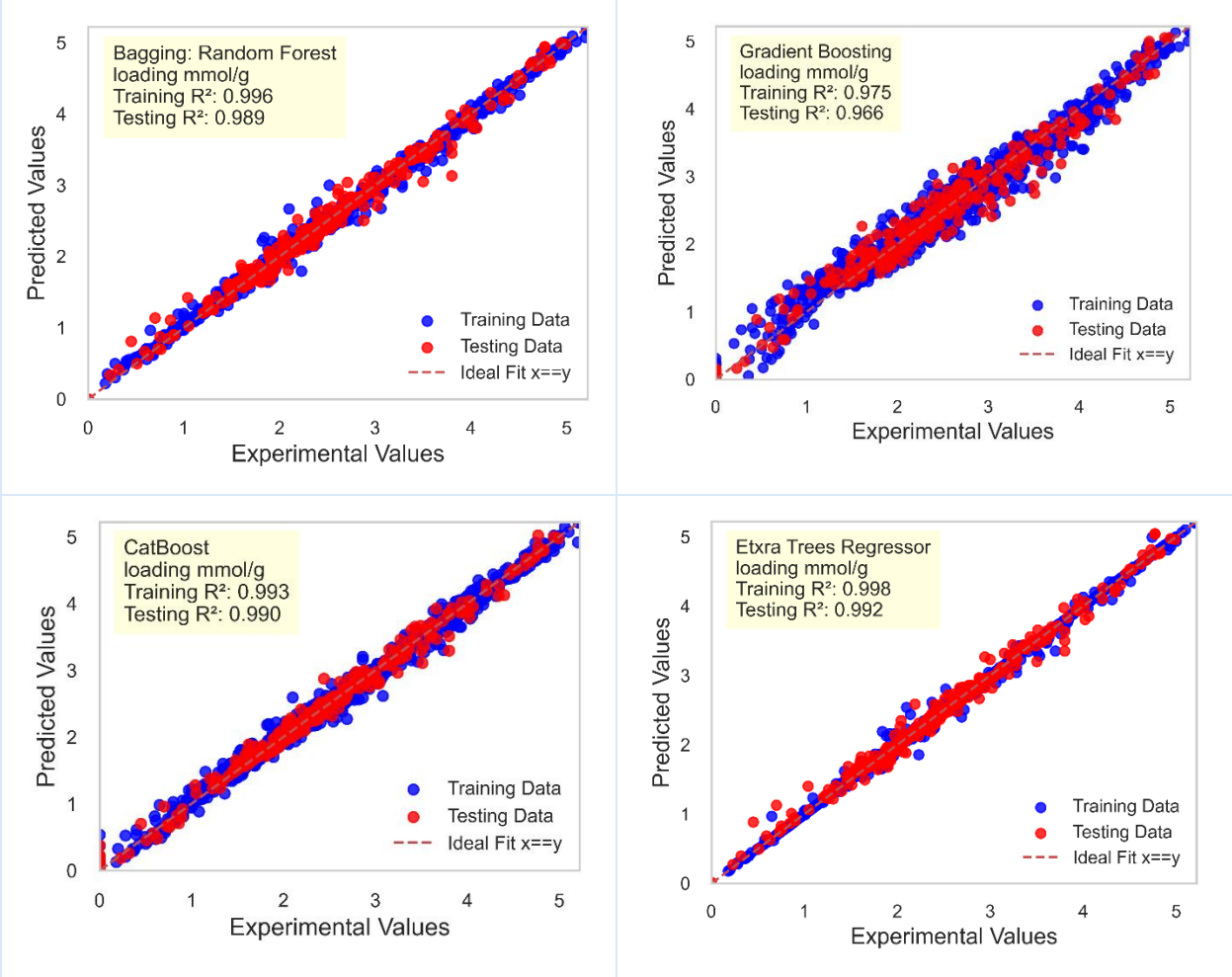


Figure 9. Original row-random parity plots for adsorption loading using Random Forest, Gradient Boosting, CatBoost, and Extra Trees. The retained figure labels use "Experimental Values"; in the revised manuscript these axes represent simulation-derived reference values. Final leakage-aware loading holdout performance is reported in the text.

4.2.3 ML Prediction of Adsorption Loading (mmol/g)

Table 5 and Figure 9 likewise retain the original row-random loading benchmark, where near-perfect R^2 values primarily reflect interpolation within the densely sampled simulation grid. In the revised analysis, Histogram Gradient Boosting was selected from the development pool. Its five-fold isotherm-grouped development performance was $R^2 = 0.974 \pm 0.007$, RMSE = 0.176 ± 0.027 mmol g^{-1}, and MAE = 0.133 ± 0.017 mmol g^{-1}. On the balanced HFC–zeolite-pair holdout, the final loading model achieved $R^2 = 0.845$, RMSE = 0.512 mmol g^{-1}, and MAE = 0.371 mmol g^{-1}. Thus, the leakage-aware evaluation remains strong but is appropriately less optimistic than the row-random benchmark.

Table 5. Original row-random interpolation benchmark for adsorption-loading models. These values are retained for comparison with the leakage-aware results reported in the text.

Model	Phase	R2	MSE	MAE	RAE
Bagging: Random Forest	Training	1.000 ± 0.000	0.002 ± 0.000	0.027 ± 0.001	0.013 ± 0.000
	Validation	0.998 ± 0.000	0.014 ± 0.004	0.070 ± 0.005	0.035 ± 0.002
	Testing	**0.999**	**0.011**	**0.059**	**0.030**
Gradient Boosting	Training	0.989 ± 0.001	0.091 ± 0.009	0.216 ± 0.014	0.109 ± 0.007
	Validation	0.986 ± 0.001	0.114 ± 0.015	0.239 ± 0.014	0.121 ± 0.008
	Testing	**0.987**	**0.112**	**0.252**	**0.129**
Advanced: CatBoost	Training	0.999 ± 0.000	0.012 ± 0.001	0.077 ± 0.004	0.039 ± 0.002
	Validation	0.997 ± 0.001	0.024 ± 0.009	0.102 ± 0.013	0.051 ± 0.005
	Testing	**0.996**	**0.036**	**0.105**	**0.054**
Extra Tress Regressor	Training	1.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Validation	0.999 ± 0.000	0.010 ± 0.003	0.058 ± 0.007	0.029 ± 0.002
	Testing	**0.999**	**0.007**	**0.050**	**0.025**

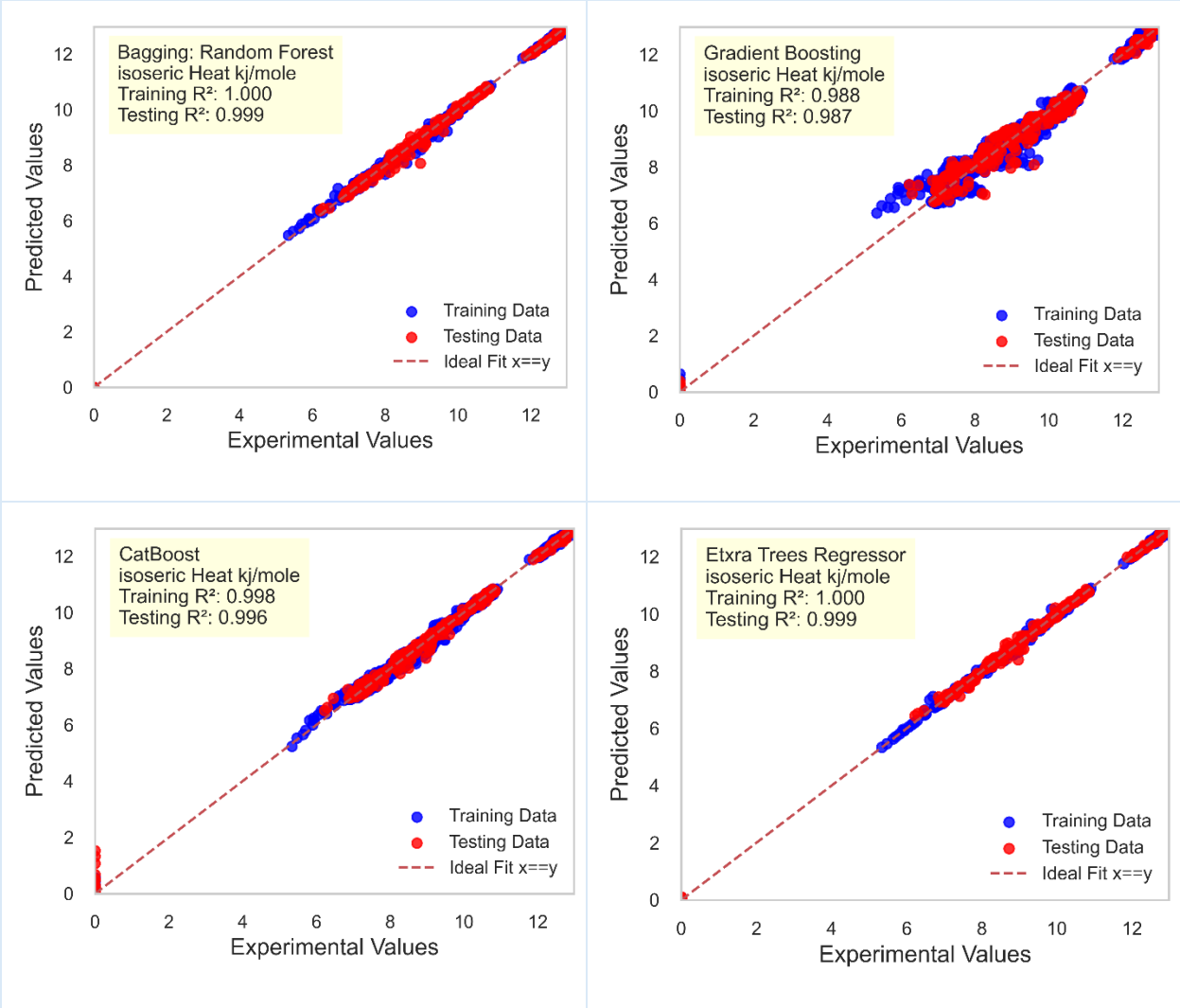


Figure 10. Original row-random parity plots for isosteric heat using Random Forest, Gradient Boosting, CatBoost, and Extra Trees. The retained figure labels use "Experimental Values"; in the revised manuscript these axes represent simulation-derived reference values. Final leakage-aware heat holdout performance is reported in the text.

4.3 Feature Optimization and Model Performance

Feature interpretation was repeated under the revised leakage-aware models and is treated as predictive attribution rather than causal physical attribution. The original SHAP figure is retained to preserve the manuscript's figure set and to document the row-random exploratory analysis. In the revised loading model, pressure had the largest mean absolute attribution, followed by the dipole-linked molecular-identity descriptor and temperature. In the revised heat model, framework density had the largest global attribution, followed by H count, VSA, void fraction, and temperature. Because these descriptors are strongly correlated within only four HFCs and four zeolites, rankings can redistribute among correlated alternatives and should be interpreted at the descriptor-family level.

The revised attribution is physically consistent with the operating trends in the simulation database: increasing pressure increases occupancy and therefore loading, whereas increasing temperature reduces loading over the sampled range. Heat prediction is more strongly tied to the local framework and molecular-composition fingerprints than loading. Nevertheless, SHAP values explain how the fitted model distributes its predictions; they do not prove that a descriptor is an independent mechanistic cause.

The contrast between the original row-random SHAP ranking in Figure 11 and the leakage-aware attribution is itself informative. A model that has access to closely neighboring rows can rely heavily on descriptors that identify a familiar system, whereas transfer to a withheld HFC–zeolite combination exposes pair-specific interaction effects. The revised Supporting Information therefore provides the full leakage-aware SHAP, permutation-importance, residual, and leave-one-identity diagnostics.

An exploratory reduced-descriptor model was also tested using seven descriptors. For loading, R^2 changed only from 0.845 to 0.838 and RMSE from 0.512 to 0.523 mmol g^{-1} on the same pair holdout. For heat, however, R^2 decreased from 0.885 to 0.711 and RMSE increased from 0.617 to 0.978 kJ mol^{-1}. Thus, a compact descriptor set may be adequate for loading interpolation, whereas heat requires a richer representation of the framework and guest environment. Because the reduced-set comparison uses the same holdout, it is exploratory rather than an additional independent validation.

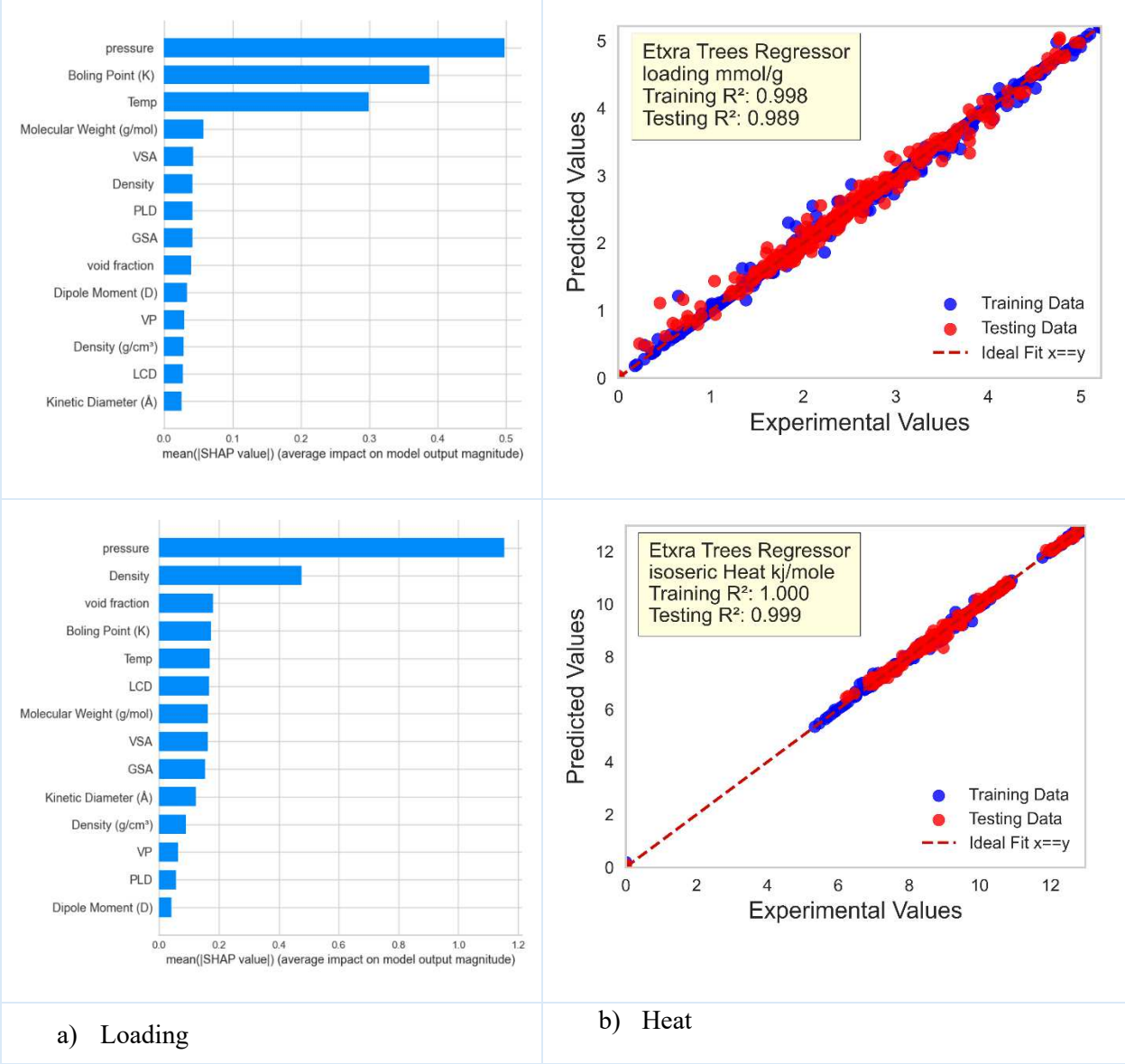


Figure 11. Original row-random SHAP feature-importance analysis retained for continuity. Revised leakage-aware attribution is discussed in the text and Supporting Information; SHAP values are interpreted as model attribution, not causal physical effects.

Generalization tests further define the applicability domain. Leave-one-HFC-out loading prediction averaged $R^2 = -0.106 \pm 0.988$, while leave-one-HFC-out heat prediction averaged $R^2 = 0.236 \pm 0.838$. Leave-one-zeolite-out loading was less severe on average ($R^2 = 0.723 \pm 0.302$), but heat extrapolation to an unseen zeolite identity was unreliable (mean $R^2 = -3.449 \pm 4.718$). These results show that the model is not a universal screening tool for arbitrary HFCs or zeolites.

Accordingly, the strongest supported use of the model is rapid interpolation over the sampled pressure–temperature domain and prediction for new combinations among the studied HFC and zeolite identities

when both component identities have been represented during development. Predictions for a new refrigerant chemistry, a new zeolite topology, or conditions substantially outside 293–333 K and 0–100 kPa require additional simulation or experimental data.

These validation and attribution results explain why the revised pair-holdout R^2 values are lower than the original row-random values while providing stronger scientific evidence. The model has not become less accurate; rather, it is being evaluated on a more demanding task that better reflects transfer beyond neighboring simulation points.

5. Conclusion and Future Work

This study preserves the original molecular-simulation adsorption database and manuscript figure set while revising the data audit and machine-learning validation to provide a more defensible estimate of predictive performance. The primary analysis contains 1,536 adsorption-loading observations and 1,440 modeled isosteric-heat values for R41, R32, R23, and R14 in 13X, 3A, 4A, and 5A across 293–333 K and 0–100 kPa. The 96 P = 0 source heat zeros are preserved in the audit data but are not treated as evaluated heat targets. Random row-level splitting reproduces the very high interpolation accuracy shown in the original parity figures, whereas isotherm-grouped development and a balanced HFC–zeolite-pair holdout provide a more stringent transfer test. Histogram Gradient Boosting achieved $R^2 = 0.845$ and RMSE = 0.512 mmol g^{-1} for loading, while Extra Trees achieved $R^2 = 0.885$ and RMSE = 0.617 kJ mol^{-1} for heat on the balanced pair holdout. Model attribution identifies pressure and temperature as key loading-related variables and correlated framework/molecular descriptor families as important for heat, but these attributions are not interpreted causally. Leave-one-HFC-out and leave-one-zeolite-out tests show that extrapolation to unseen identities remains unreliable. The resulting model is therefore best described as a transparent interpolation surrogate for the studied HFC–zeolite domain. Future work should expand the chemical diversity, confirm the source descriptor definitions, incorporate additional experimentally validated systems, and evaluate true external generalization.

Data and Software Availability

The machine-readable audited dataset, split assignments, source-integrity audit, model outputs, and reproducibility scripts are provided with this submission. The raw adsorption targets are preserved numerically and are traceable to the supplied source workbook and the previously published molecular-simulation study. The undocumented second 1,536-row extension is preserved separately but is not used in the primary inferential analysis because explicit HFC identities and complete descriptor provenance are not available in the supplied source files.

Acknowledgment

The authors gratefully acknowledge the support provided by King Fahd University of Petroleum and Minerals (KFUPM), where this research work was conducted. The authors thank Dr. Hamdi Ali Ahmed Al-Jamimi from the Information and Computer Science Department, College of Computing and Mathematics, for guidance on the original machine-learning study. The authors also acknowledge the computational and research facilities available at KFUPM. The molecular-simulation targets reused in this analysis originate from the previously published study cited in the manuscript.

Funding

No funding was received for conducting this study.

References

Abdelrasoul A, Zhang H, Cheng C-H, Doan H (2017) Applications of molecular simulations for separation and adsorption in zeolites. Microporous Mesoporous Mater 242:294–348.

Achour S, Hosni Z (2024) ML-driven models for predicting CO2 uptake in metal–organic frameworks (MOFs). Can J Chem Eng

Al-Sakkari EG, Ragab A, So TMY, Shokrollahi M, Dagdougui H, Navarri P et al (2023) Machine learning-assisted selection of adsorption-based carbon dioxide capture materials. J Environ Chem Eng 11(5):110732.

Baker RW, Lokhandwala K (2008) Natural gas processing with membranes: an overview. Ind Eng Chem Res 47(7):2109–2121.

Barpaga D, Zheng J, McGrail BP, Motkuri RK (2021) Manipulating pore topology and functionality to promote fluorocarbon-based adsorption cooling. Acc Chem Res 55(5):649–659.

Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A (2018) Machine learning for molecular and materials science. Nature 559(7715):547–555.

Castle WF (2002) Air separation and liquefaction: recent developments and prospects for the beginning of the new millennium. Int J Refrig 25(1):158–172.

Cejka J, Corma A, Zones S (2010) Zeolites and catalysis: synthesis, reactions and applications. John Wiley & Sons.

Chakraborty A, Gandhi A, Hasan MMF, Venkatasubramanian V (2024) Discovering zeolite adsorption isotherms: a hybrid AI modeling approach. In: Computer Aided Chemical Engineering. Vol 53. Elsevier. pp 511–516.

De Monte F (2002) Calculation of thermodynamic properties of R407C and R410A by the Martin–Hou equation of state—part II: technical interpretation. Int J Refrig 25(3):314–329.

Elhussien AA, Abdulazeez I, Alasiri H, Fouad WA (2025) Exploring zeolite potential for hydrofluorocarbon capture and recycling: Insights from molecular simulations. Microporous Mesoporous Mater 384(December 2024):113442. https://doi.org/10.1016/j.micromeso.2024.113442

Evans JD, Coudert F-X (2024) Challenges and opportunities of molecular simulations for negative gas adsorption. Acc Mater Res 5(5):640–647.

Fang M, Liu F, Wang T, Zhu D, Dong W, Xu Y et al (2022) Chemical absorption. In: Handbook of Climate Change Mitigation and Adaptation. Springer. pp 1405–1481.

Fanourgakis GS, Gkagkas K, Tylianakis E, Froudakis G (2020) A generic machine learning algorithm for the prediction of gas adsorption in nanoporous materials. J Phys Chem C 124(13):7117–7126.

Freni A, Dawoud B, Bonaccorsi L, Chmielewski S, Frazzica A, Calabrese L et al (2015) Characterization of zeolite-based coatings for adsorption heat pumps. Springer.

Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR et al (2016) Gaussian 16 Rev. C.01. Wallingford.

Gupta TK, Raza K (2019) Optimization of ANN architecture: a review on nature-inspired techniques. Machine learning in bio-signal analysis and diagnostic imaging 159–182.

Khaleque A, Alam MM, Hoque M, Mondal S, Bin Haider J, Xu B et al (2020) Zeolite synthesis from low-cost materials and environmental applications: A review. Environ Adv 2:100019.

Klepov VV, Juillerat CA, Alekseev EV, Zur Loye H-C (2018) Overstepping Löwenstein's rule—a route to unique aluminophosphate frameworks with three-dimensional salt-inclusion and ion-exchange properties. Inorg Chem 58(1):724–736.

Li C, Bao L, Ji Y, Tian Z, Cui M, Shi Y et al (2024) Combining machine learning and metal–organic frameworks research: Novel modeling, performance prediction, and materials discovery. Coord Chem Rev 514:215888.

Li J-R, Kuppler RJ, Zhou H-C (2009) Selective gas adsorption and separation in metal–organic frameworks. Chem Soc Rev 38(5):1477–1504.

Luo J, Quan Y, Xu S (2023) Robust-GBDT: A Novel Gradient Boosting Model for Noise-Robust Classification. arXiv preprint arXiv:2310.05067

Ma Y, Zhang F, Lively RP (2020) Manufacturing nanoporous materials for energy-efficient separations: Application and challenges. In: Sustainable Nanoscale Engineering. Elsevier. pp 33–81.

Maiorino A, Aprea C, Del Duca MG, Llopis R, Sánchez D, Cabello R (2018) R-152a as an alternative refrigerant to R-134a in domestic refrigerators: an experimental analysis. Int J Refrig 96:106–116. https://doi.org/10.1016/j.ijrefrig.2018.09.020

Materials Studio (2017) Materials Studio, revision 8.0. Accelrys Inc., San Diego, USA.

Mollahosseini A, Abdelrasoul A (2021) Molecular dynamics simulation for membrane separation and porous materials: A current state of art review. J Mol Graph Model 107:107947.

Myhre G, Shindell D, Bréon F-M, Collins W, Fuglestvedt J, Huang J et al (2014) Anthropogenic and natural radiative forcing. Climate Change 2013-The Physical Science Basis 659–740.

Oliveira LMC, Koivisto H, Iwakiri IGI, Loureiro JM, Ribeiro AM, Nogueira IBR (2020) Modelling of a pressure swing adsorption unit by deep learning and artificial Intelligence tools. Chem Eng Sci 224:115801.

Purohit P, Höglund-Isaksson L (2017) Global emissions of fluorinated greenhouse gases 2005–2050 with abatement potentials and costs. Atmos Chem Phys 17(4):2795–2816. https://doi.org/10.5194/acp-17-2795-2017

Pérez-Botella E, Valencia S, Rey F (2022) Zeolites in adsorption processes: State of the art and future prospects. Chem Rev 122(24):17647–17695.

Ravishankara AR, Daniel JS, Portmann RW (2009) Nitrous oxide (N2O): the dominant ozone-depleting substance emitted in the 21st century. Science 326(5949):123–125.

Russell AJ, Spackman MA (2000) An ab initio study of vibrational corrections to the electrical properties of the fluoromethanes: CH3F, CH2F2, CHF3 and CF4. Mol Phys 98(10):633–642. https://doi.org/10.1080/00268970009483330

Sadeghi P, Myers MB, Pareek V, Arami-Niya A (2024) Temperature-regulated gas adsorption on LTA zeolites: Observation of sorption isotherms for methane, hydrogen, nitrogen and carbon dioxide. Appl Surf Sci 678:161037.

Sagi O, Rokach L (2021) Approximating XGBoost with an interpretable decision tree. Inf Sci 572:522–542.

Salahshoori I, Mahdavi S, Moradi Z, Otadi M, Kazemabadi FZ, Nobre MAL et al (2024a) Advancements in molecular simulation for understanding pharmaceutical pollutant Adsorption: A State-of-the-Art review. J Mol Liq 125513.

Salahshoori I, Wang Q, Nobre MAL, Mohammadi AH, Dawi EA, Khonakdar HA (2024b) Molecular simulation-based insights into dye pollutant adsorption: a perspective review. Adv Colloid Interface Sci 103281.

Smit B, Maesen TLM (2008) Molecular simulations of zeolites: adsorption, diffusion, and shape selectivity. Chem Rev 108(10):4125–4184.

Sun Y, Ma Y, Yu B, Zhang W, Zhang L, Chen P et al (2024) Perspectives on Molecular Simulation of CO2/CH4 Competitive Adsorption in a Shale Matrix: A Review. Energy Fuels 38(17):15935–15971.

Taddei M, Petit C (2021) Engineering metal–organic frameworks for adsorption-based gas separations: From process to atomic scale. Mol Syst Des Eng 6(11):841–875.

Thommes M, Cychosz KA (2014) Physical adsorption characterization of nanoporous materials: progress and challenges. Adsorption 20(2):233–250.

Wang M, Lawal A, Stephenson P, Sidders J, Ramshaw C (2011) Post-combustion CO2 capture with chemical absorption: A state-of-the-art review. Chem Eng Res Des 89(9):1609–1624.

Wang C, Zhao Y, Gao Y, Chen H, Li X, Zhou B et al (2024) Interpretable machine learning for predicting heavy metal removal and optimizing biochar characteristics. J Water Process Eng 68:106484.

Wanigarathna JADK (2018) Adsorption-based fluorocarbon separation in zeolites and metal organic frameworks.

Xiang HW (2001) Vapor pressures, critical parameters, boiling points, and triple points of halomethane molecular substances. J Phys Chem Ref Data 30(5):1161–1197. https://doi.org/10.1063/1.1403734

Yan Y, Zhang L, Li S, Liang H, Qiao Z (2021) Adsorption behavior of metal-organic frameworks: From single simulation, high-throughput computational screening to machine learning. Comput Mater Sci 193:110383.

Yancey AD, Corbin DR, Shiflett MB (2022a) Difluoromethane (HFC-32) and Pentafluoroethane (HFC-125) Sorption on Linde Type A (LTA) Zeolites for the Separation of Azeotropic Hydrofluorocarbon Refrigerant Mixtures. Langmuir 38(6):1937–1953. https://doi.org/10.1021/acs.langmuir.1c02904

Yancey AD, Terian SJ, Shaw BJ, Bish TM, Corbin DR, Shiflett MB (2022b) A review of fluorocarbon sorption on porous materials. Microporous Mesoporous Mater 331:111654.

Yang X, Yuan C, He S, Jiang D, Cao B, Wang S (2023) Machine learning prediction of specific capacitance in biomass derived carbon materials: Effects of activation and biochar characteristics. Fuel 331:125718.

Yulia F, Chairina I, Zulys A (2021) Multi-objective genetic algorithm optimization with an artificial neural network for CO2/CH4 adsorption prediction in metal–organic framework. Therm Sci Eng Prog 25:100967.

Zhong S, Zhang K, Bagheri M, Burken JG, Gu A, Li B et al (2021) Machine learning: new ideas and tools in environmental science and engineering. Environ Sci Technol 55(19):12741–12754.