Supplementary Information

Title

AI-driven discovery of high-performance metal-organic frameworks for next-gen atmospheric water harvesting

Authors

Samar Abdelwadood[1*], Seda Rouxel[2], Samuel Mao[1*] and Ludovic F. Dumée[3,4*]

Affiliations

[1] Department of Mechanical and Nuclear Engineering, Khalifa University, Abu Dhabi, UAE

[2] Department of Earth Sciences, Khalifa University, Abu Dhabi, UAE

[3] Element Zero, Research and Innovation Department, Malaga, Western Australia, Australia

[4] Nanjing Tech University, Nanjing, China

Corresponding authors

*samar.elwadood@ku.ac.ae; samuel.mao@ku.ac.ae; ludo@elementzero.green

S1 Methodology

S1.1 Dataset Curation and Processing

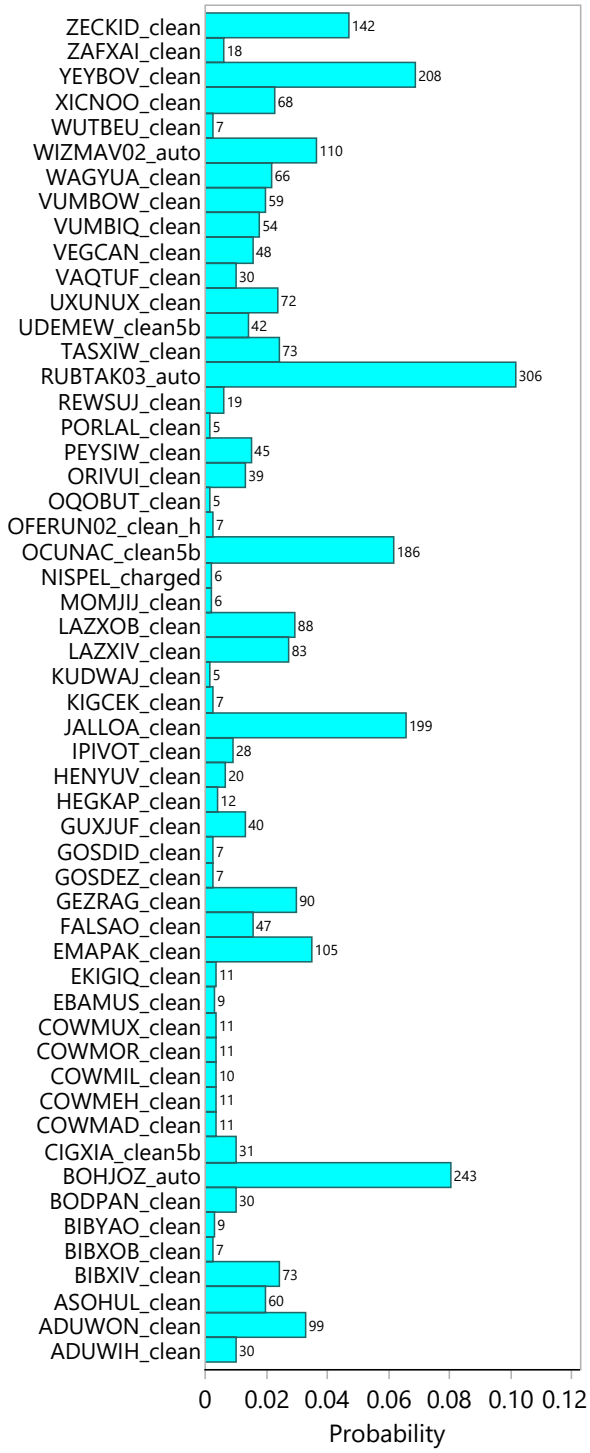

Figure S1 Distribution of MOFs in the experimental dataset. This histogram illustrates the frequency and relative representation of each MOF based on the number of experimental measurements available in the compiled dataset.

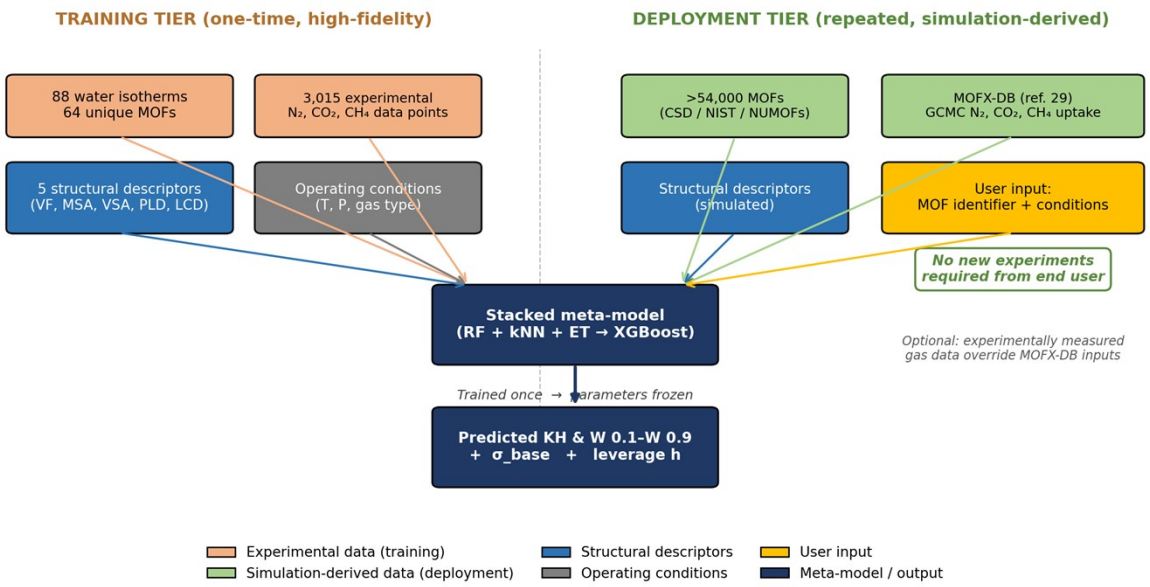

Figure S2 Two-tier hybrid descriptor architecture of the meta-model. Training tier (left, peach): the meta-model is fitted to 88 experimental water-adsorption isotherms across 64 MOFs together with 3,015 experimental N_2, CO_2 and CH_4 data points. The training is performed once and the parameters are frozen thereafter. Deployment tier (right, green): the open-access screening tool operates on >54,000 MOFs whose gas-adsorption descriptors are obtained from MOFX-DB and analogous simulation-derived repositories. The end user supplies only a MOF identifier and the operating conditions; if experimental gas-adsorption data are available for the candidate, they may be entered directly to override the simulation-derived defaults. The same trained meta-model serves both tiers and outputs the predicted Henry's constant and water-uptake vector together with the base-learner.

S1.2 Dimensionality Reduction Techniques

Dimensionality reduction techniques, specifically Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE), were applied to identify patterns and trends hidden within the data set. PCA transformed the original features into orthogonal principal components to retain the highest amount of variability. PCA enabled the identification of the most important features and created a simplified feature space to enable further analysis (*1*). t-SNE enabled a non-linear visual representation of the high dimensional data to aid in the discovery of clusters and trends in structural and adsorption properties of MOFs (*2*). These techniques provided insight into the clustering and variability of the MOF property data to allow for the identification of patterns crucial for predictive modeling.

Summary statistics, distribution plots and box plots were used to provide an overview of the variability and center of the distribution of the collected and screened data sets and thus provide a thorough understanding of the distribution of the features.

S1.3 Pearson Correlation Coefficient

The Pearson correlation coefficient (PCC) was used to determine the linear relationship between the structural descriptors and the adsorption properties being targeted. The PCC varies from -1 to 1, where numbers close to 1 or -1 signify a strong positive or negative correlation, respectively, and numbers close to 0 signify little to no correlation. Features that have a high correlation value with the target variables will be considered first for training machine learning models, while features with a high correlation value among themselves will be considered for further investigation. This will help to ensure that the features most responsible for the adsorption properties are being trained upon, thereby increasing the predictive capability and interpretability of the models (*3*).

S1.4 Decision Tree Feature Importance

Decision tree feature importance capabilities within JMP SAS Pro (version 16) were also used to evaluate feature importance. The importance of each feature was determined by calculating how much each feature contributed to the reduction of uncertainty in the model output at each node in the tree. For any feature 'f', the importance of the feature 'I(f)' is calculated as the sum of the reduction of the error in predicting the sample that reached each node that split on 'f' times the fraction of the total sample that reached each node. When ensemble methods, such as Random Forest, are used, then the average of the feature importance scores from the individual trees are used to create an overall score for each feature. This allows for the determination of the most important features in the model, aids in the interpretation of the data, and helps guide the improvement of the modeling process (*4*).

S1.5 Machine Learning Framework

A robust machine learning framework was developed to predict water adsorption properties and reconstruct complete water adsorption isotherms. This framework employed a diverse array of algorithms, each tailored to capture specific aspects of the data's structural and adsorption complexity. Linear models such as Multiple Linear Regression (MLR), Ridge Regression, and Lasso Regression served as baselines, providing insights into the linear relationships between structural descriptors and adsorption properties. Non-linear models, including k-Nearest Neighbours (kNN), Random Forest (RF), Extra Trees (ET), Adaptive Boosting (AdaBoost), Extreme Gradient Boosting (XGB), Gradient Boosting (GB), Light Gradient Boosting Machine (LGBM), and Artificial Neural Networks (ANNs), were employed to capture intricate, non-linear dependencies.

The framework integrated structural descriptors such as void fraction, MSA, VSA, PLD, and LCD with adsorption metrics like Henry's constant and water uptake values. Each algorithm underwent rigorous hyperparameter tuning using random search, grid search, and Bayesian optimization. Cross-validation and splitting the dataset 80/20% for training and testing, respectively, were employed to prevent overfitting and ensure the generalizability of the models. Model performance was evaluated using key metrics, including the coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE), calculated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_{i,exp} - y_{i,pred}\right| \qquad \textit{Equation S1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{i,exp} - y_{i,pred}\right)^2} \qquad \textit{Equation S2}$$

Where $y_{i,exp}$ and $y_{i,pred}$ represent the experimental and predicted values, respectively, and n denotes the number of data points. These metrics provided a comprehensive evaluation of the models' predictive accuracy and error characteristics (*5*).

S1.6 Meta-Model Development

The concept of meta-modelling, also known as model stacking or stacked generalization, represents an advanced ensemble learning approach that combines predictions from multiple base models through a higher-level model to improve predictive accuracy. First introduced by Wolpert (*6*), meta-modelling has emerged as a powerful technique for handling complex prediction tasks by leveraging the complementary strengths of different algorithms. In traditional ensemble methods, such as bagging or boosting, predictions are typically combined through simple averaging or weighted voting. However, meta-modelling takes a more sophisticated approach by treating the outputs of base models as features for a secondary model, which learns optimal combination strategies that can capture non-linear relationships between different model predictions (*7*).

The theoretical foundation of meta-modelling rests on the concept of bias-variance decomposition in machine learning. Each base model exhibits different bias-variance characteristics based on its underlying assumptions and architecture. The framework can potentially achieve lower overall prediction error than any individual model by combining these models through a meta-learner. This error reduction occurs because the meta-learner can learn to compensate for individual model weaknesses while amplifying their strengths across different regions of the feature space.

In our implementation, the meta-model architecture consists of three levels:

1. Base Model Layer: Comprises three diverse algorithms (Random Forest, k-Nearest Neighbours, and Extra Trees) chosen for their complementary learning approaches.

Random Forest provides robust ensemble predictions through bagging, k-Nearest Neighbours captures local patterns in the feature space, and Extra Trees introduces beneficial randomization in the splitting process. The Random Forest model achieved particularly notable results, with training R^2 values ranging from 0.995 for KH (Henry's constant) to 0.998 for working capacity at high relative humidity (W 0.4). The k-Nearest Neighbours algorithm showed consistent performance with R^2 values between 0.993 and 0.997, while Extra Trees maintained robust accuracy with R^2 values from 0.974 to 0.998.

2. Feature Engineering Layer: Transforms base model predictions into an enriched feature space that includes both direct predictions and derived features capturing prediction uncertainties and inter-model agreements (*8*).

3. Meta-Learner Layer: Employs XGBoost as the final estimator, leveraging its gradient boosting framework to learn optimal prediction combinations. The choice of XGBoost is motivated by its ability to handle non-linear relationships and its robust performance in handling heterogeneous feature spaces (*9*, *10*).

This hierarchical design enables the simultaneous prediction of all target variables, such as Henry's constant (KH) and water uptake at various relative pressures (W0.1–W0.9), leveraging ensemble learning to address the inherent weaknesses of individual models. The meta-model's architecture, built using a multi-output regression approach, enabled simultaneous prediction of all target variables, reducing computational overhead and improving performance by leveraging shared information between variables. The training process, employing an 80-20 train-test split, demonstrated exceptional performance. The integration of XGBoost as the meta-learner provided an additional layer of optimization, enabling efficient handling of complex feature spaces while maintaining computational efficiency.

The training process follows a k-fold cross-validation strategy to prevent information leakage between the base models and meta-learner. This ensures that the meta-learner receives unbiased predictions from the base models during both training and inference phases. The meta-learner's hyperparameters are optimized to maximize prediction accuracy while maintaining computational efficiency (*11*).

This meta-modelling approach offers several theoretical advantages for our multi-property prediction task:

- Enhanced Generalization: By combining diverse learning algorithms, the meta-model can better capture different aspects of the underlying structure-property relationships in MOFs (*12*).
- Reduced Overfitting: The multi-level architecture provides natural regularization, as prediction errors from individual models tend to cancel out when combined optimally (*13*).
- Improved Uncertainty Estimation: The meta-model framework allows for better quantification of prediction uncertainties by considering disagreements between base models.
- Feature Space Enrichment: The transformation of base model predictions into meta-features creates a richer representation space for final predictions.

This theoretically grounded approach to meta-modelling provides a robust framework for handling the complex, multi-property prediction challenges inherent in MOF screening for atmospheric water harvesting applications.

S1.7 Applicability Domain and Interpretability

The applicability domain (AD) delineates the region in the feature space where the model's predictions are deemed reliable. Predictions outside this domain may have low accuracy due to the high degree of extrapolation beyond the training data. The AD assessment relied on two primary metrics: leverage values and standardized residuals (SDRs). Leverage values represent the distance of a data point from the centroid of the predictor variables and were calculated using the hat matrix H:

$$H = X\left(X^T X\right)^{-1} X^T \quad \textit{Equation S3}$$

Where X is the input feature matrix with dimensions p×k (p being the number of data points and k the number of input features). The diagonal elements of H indicate the leverage value of each data point. Data points with leverage values exceeding the critical threshold h*, defined as outliers:

$$h^* = 3\frac{k+1}{p_{train}} \quad \textit{Equation S4}$$

Standardized residuals quantified the difference between the experimental and predicted values relative to the model's variability. For a data point i, the standardized residual was calculated as:

$$SDR_i = \frac{y_{i,exp} - y_{i,pred}}{\sqrt{\frac{1}{n}\sum\left(y_{i,exp} - y_{i,pred}\right)^2}} \quad \textit{Equation S5}$$

Data points with $|SDR_i| > 3$ were considered outside the AD. A Williams plot, illustrating standardized residuals against leverage values, was utilized to visualize the AD and identify outliers.

The scope of the AD was further quantified using:

$$AD_{scope} = 100 \times \frac{p_{inside}}{p} \quad \textit{Equation S6}$$

Where p_{inside} denotes the number of data points within the AD, and p is the total number of data points. A high AD_{scope} indicates that the model can reliably predict a large portion of the data space. The AD analysis provided a comprehensive evaluation of model reliability, guiding its application to new MOFs by combining leverage values and SDRs (*14*).

S1.8 SHAP Analysis

To elucidate the contribution of individual features to model predictions, the Shapley Additive Explanations (SHAP) method was employed. SHAP analysis provided an additive measure of feature contributions for each prediction, ensuring transparency in model decision-making. SHAP facilitated a deeper understanding of the structure-property relationships governing MOF performance, guiding the rational design of materials for atmospheric water harvesting (AWH) by quantifying feature importance on a per-instance basis. The SHAP analysis was conducted using the Tree Explainer method from the SHAP library in Python. This method calculates SHAP values by approximating the complex models with simpler, interpretable models (*15*). SHAP analysis enhanced our understanding of the underlying relationships between MOF descriptors and their adsorption performance by providing feature-level interpretability. This facilitates targeted material design by identifying key features to optimize for AWH.

S1.9 Uncertainty quantification

Predictive uncertainty was quantified directly from the external-validation residuals (n = 27 paired predicted–measured values across three MOFs do not present in the training set: EZOFEF from the NIST adsorption database; ADAXIO and XICNOO02 measured at 25 °C in this work). The pooled residual distribution has a sample standard deviation of s = 4.27 mmol/g and a near-zero mean (−0.42 mmol/g, indicating a small under-prediction bias). Under a Gaussian approximation, the 95 % prediction interval is therefore $\hat{y} \pm 1.96 \cdot s = \hat{y} \pm 8.37$ mmol/g; the empirical 2.5^{th} - 97.5^{th} residual percentiles give an asymmetric interval of (−8.10, +6.52) mmol/g, in close agreement with the Gaussian estimate. For per-prediction reliability the screening tool also reports σ_base, the standard deviation of the three base learners' predictions (Random Forest, k-Nearest Neighbours, Extra Trees) for the input vector of interest. σ_base captures the structural component of the uncertainty associated with finite training data and is exposed automatically alongside every meta-model output, providing an input-specific complement to the population-level prediction interval.

S2 Results and discussion

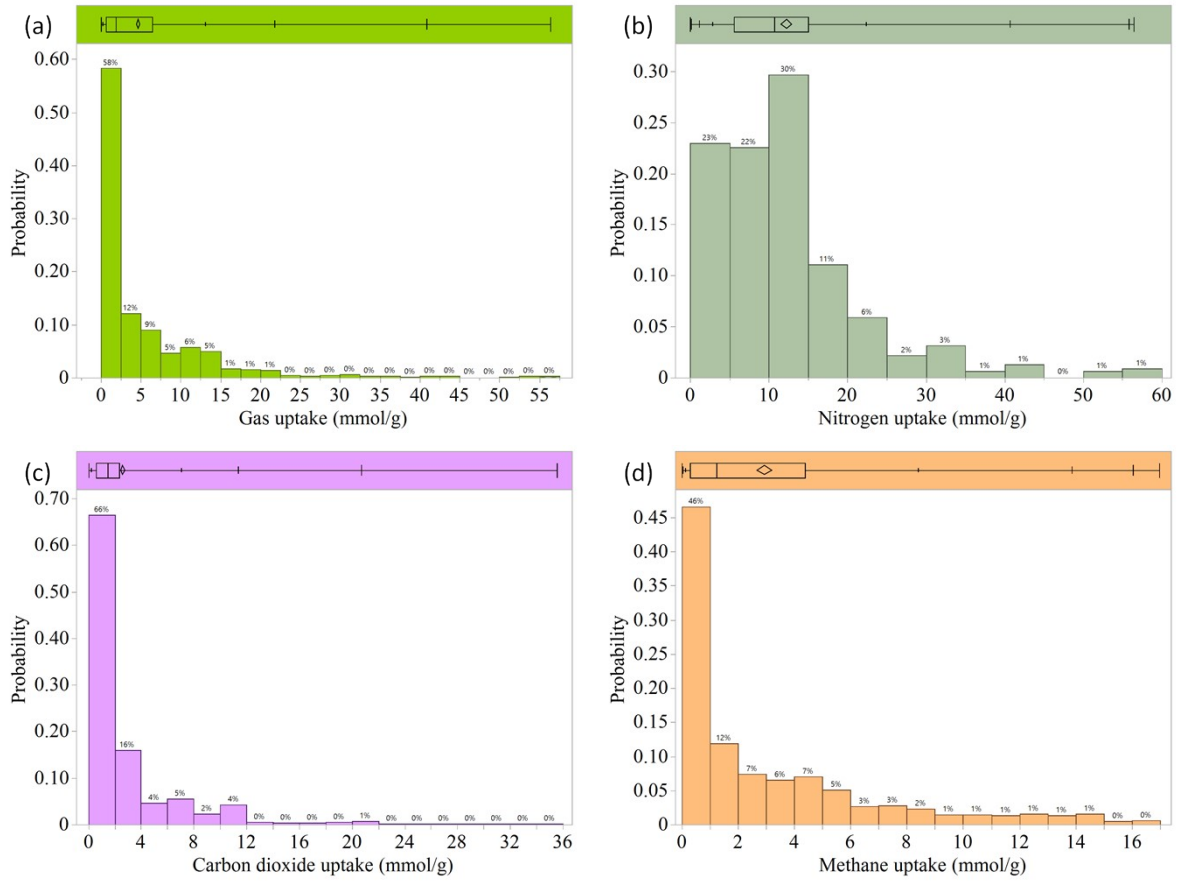

Figure S3 Visualization of gas uptake distributions for (a) all gases, (b) N_2, (c) CO_2, and (d) CH_4 data across varying pressure and temperature conditions.

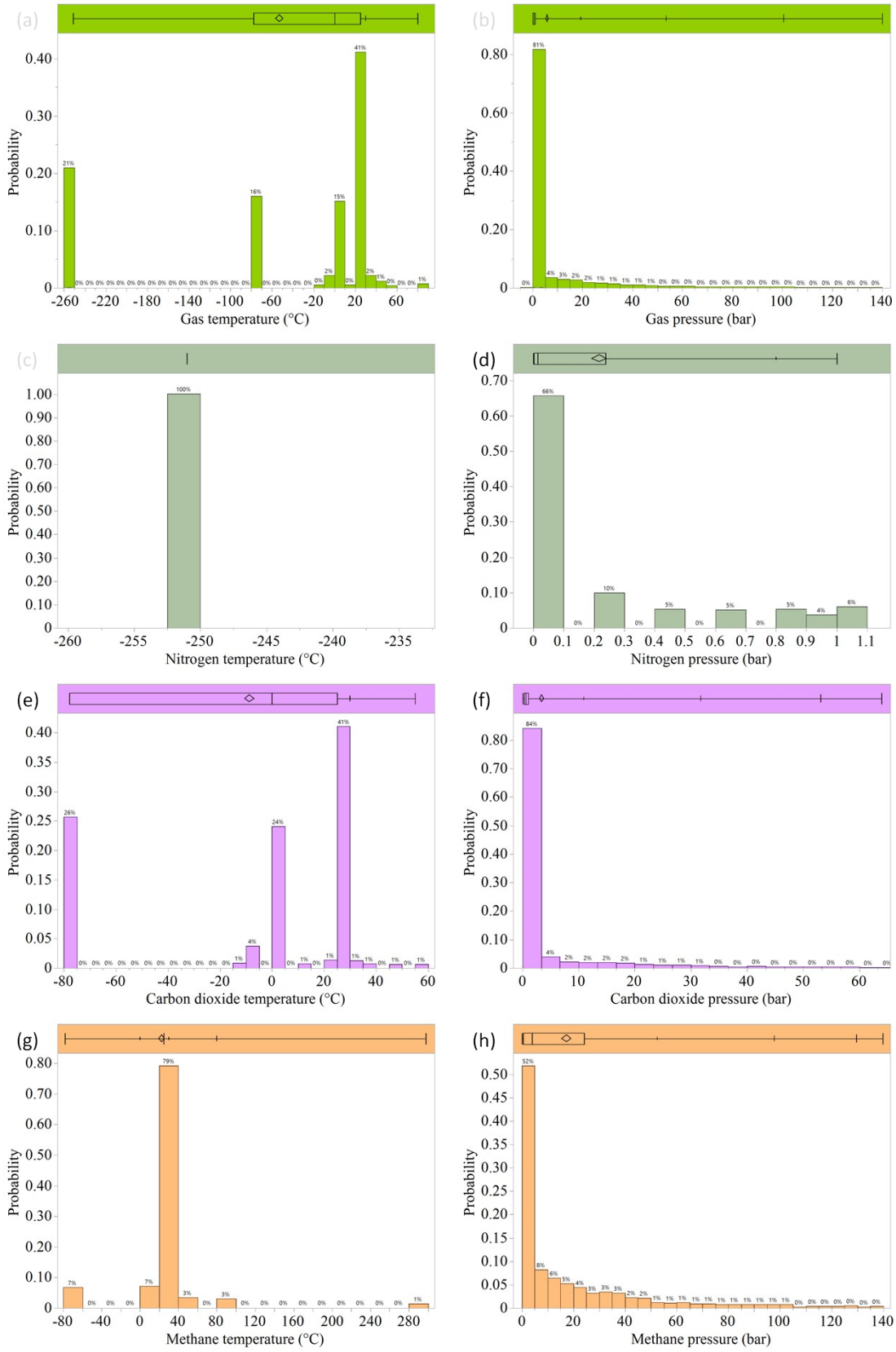

Figure S4 Distribution of gas pressures and temperature in the experimental dataset: (a, b) all gases data, (c, d) N_2 data, (e, f) CO_2 data, and (g, h) CH_4 data.

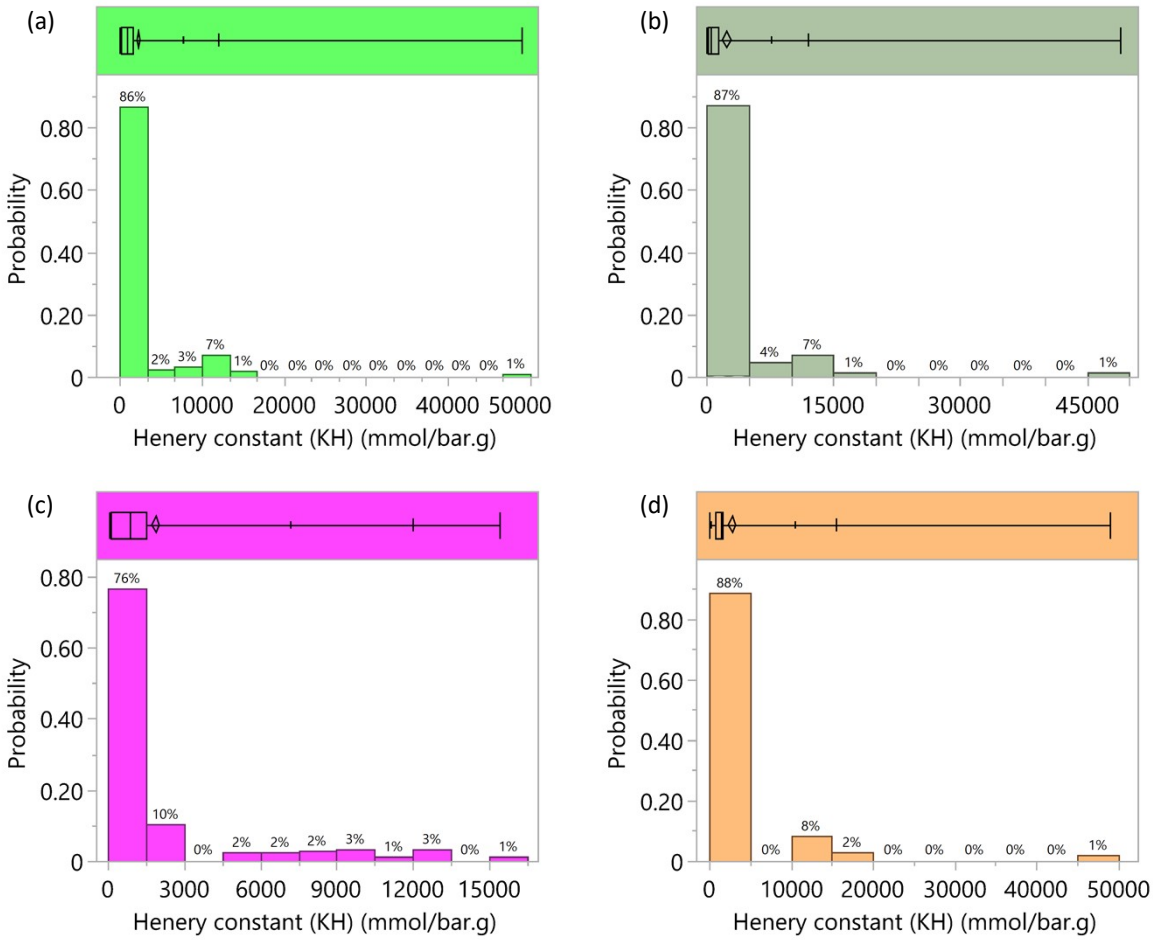

Figure S5 Henry's constant (KH) distribution for water adsorption: (a) all gases data, (b) N_2 data, (c) CO_2 data, and (d) CH_4 data.

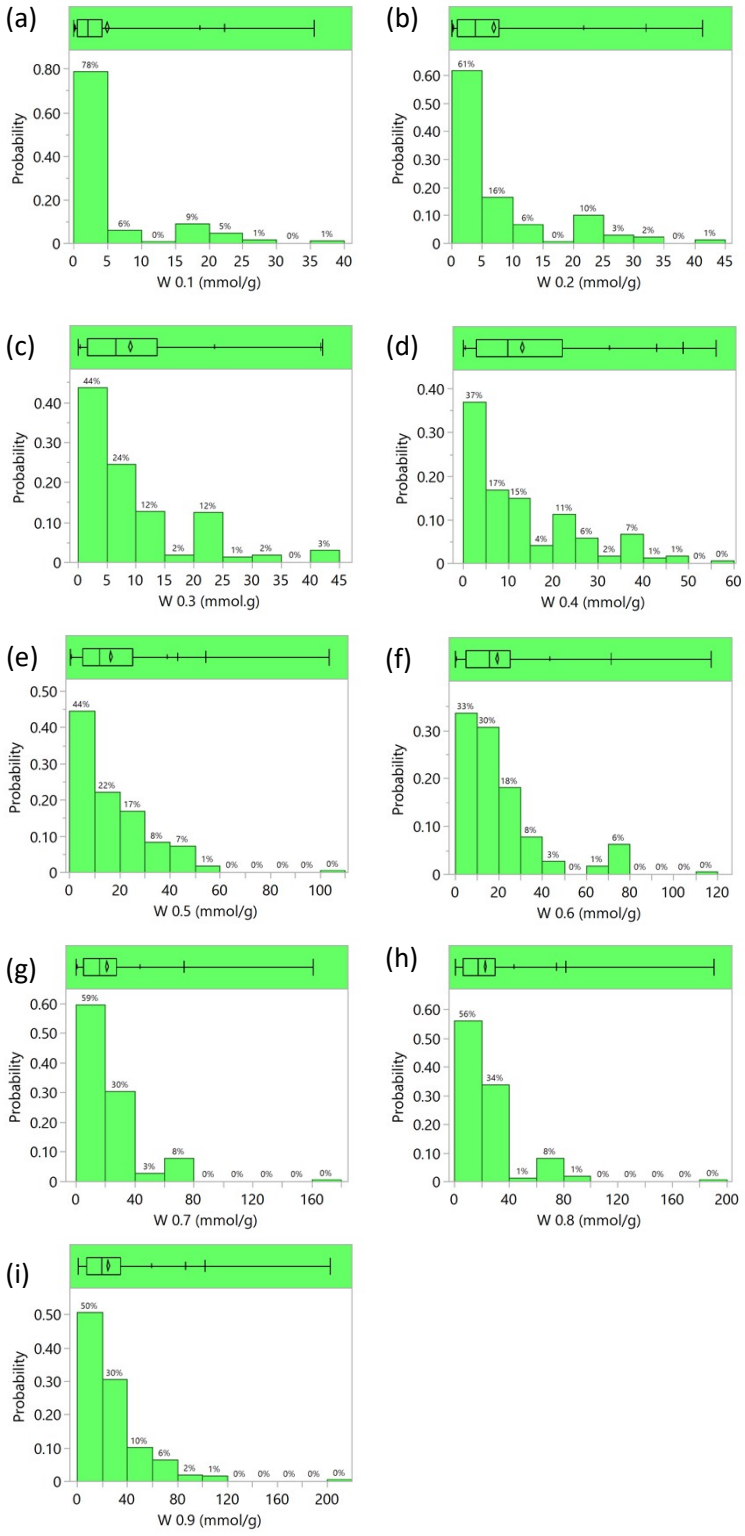

Figure S6 Distribution of water uptake values (W0.1–W0.9) at increasing relative pressures from 10% to 90% RH.

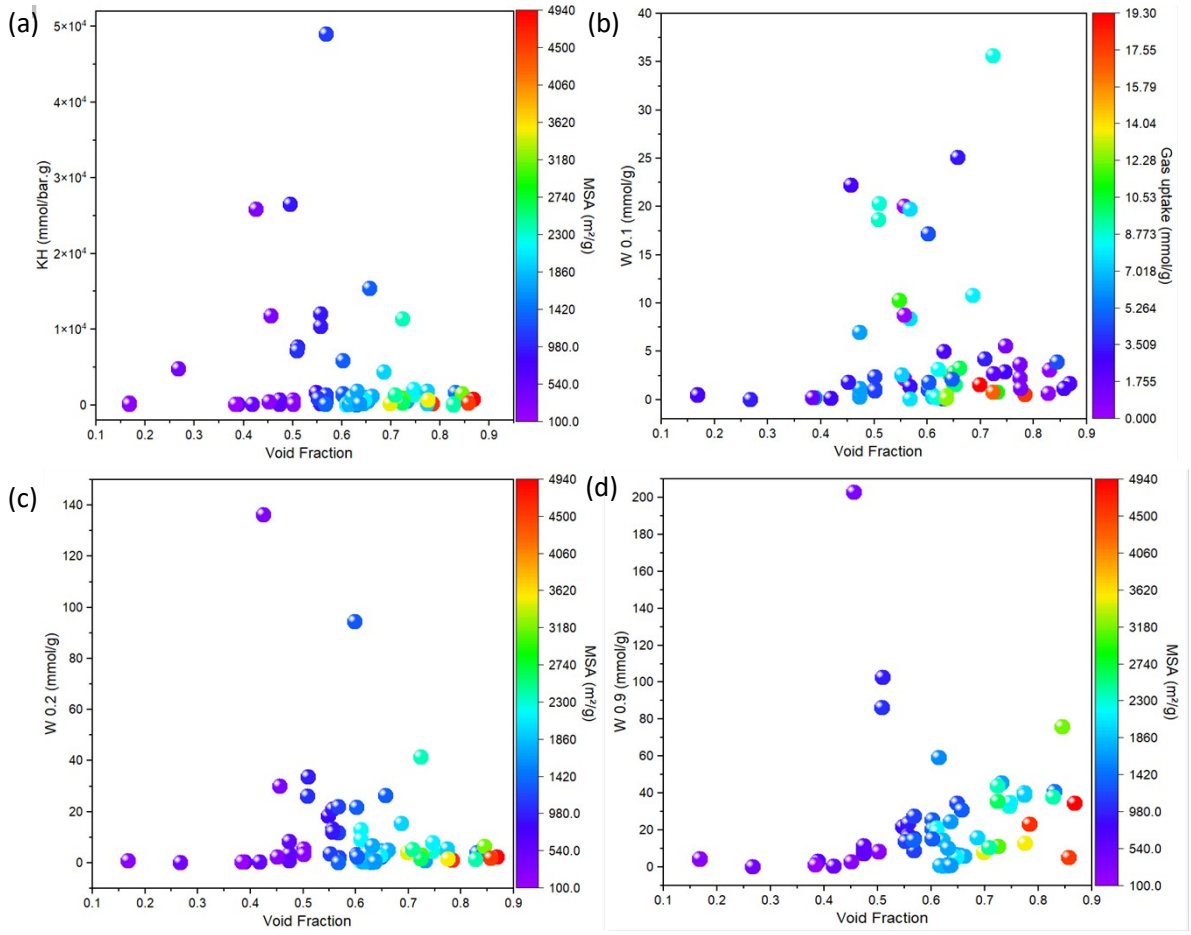

Figure S7 Bubble plots showing the correlation between key water adsorption variables and structural descriptors in the training dataset. (a) KH, (b) W0.1, (c) W0.2, and (d) W0.9 vs. void fraction. Bubble size represents the pore limiting diameter (PLD), while colour indicates MSA from low (purple) to high (red).

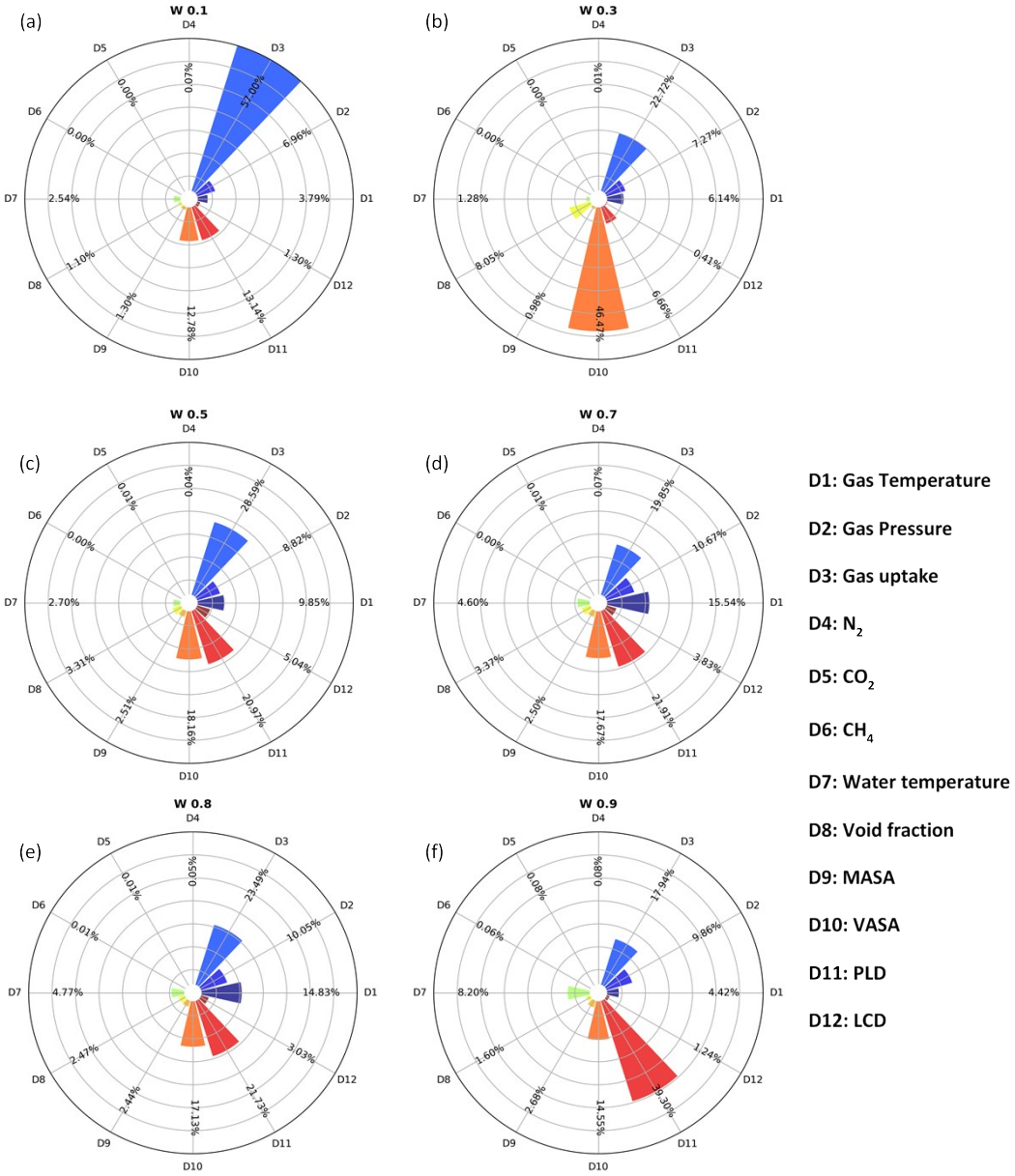

Figure S8 Radial plots illustrating the feature importance for each of the ten water adsorption properties under different weighting scenarios (KH, W0.1–W0.9). Each wedge corresponds to one of the twelve input features (D1–D12), with the wedge size indicating its relative contribution to the model.

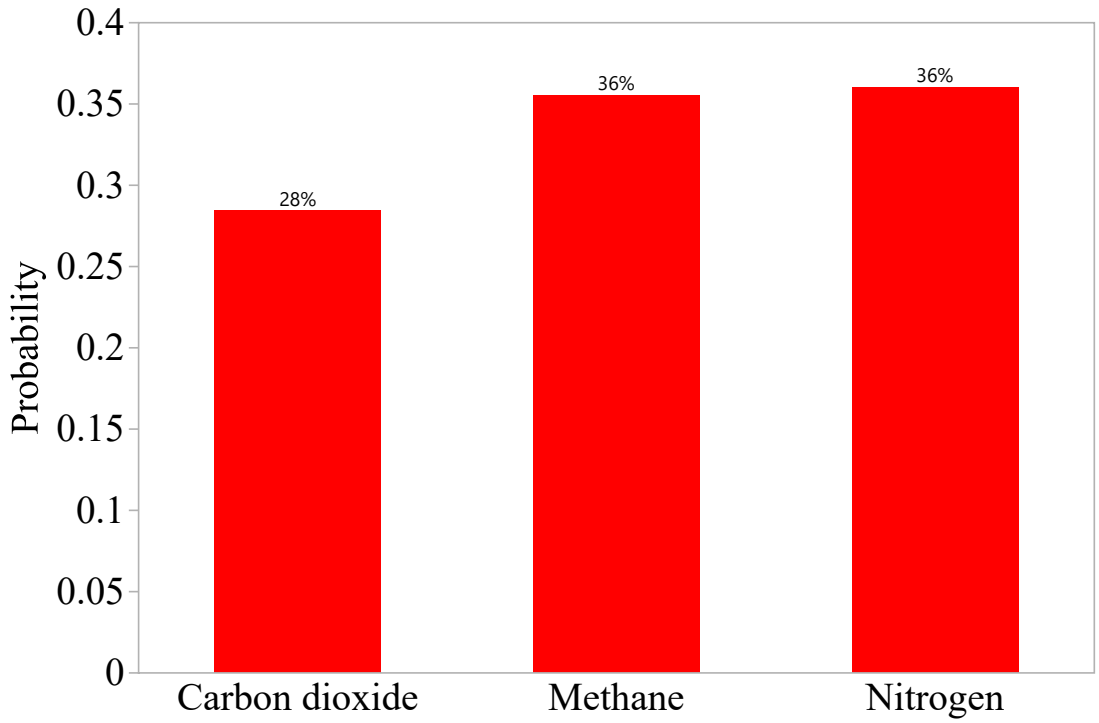

Figure S9 Bar chart showing the probability distribution of carbon dioxide, methane, and nitrogen in the screened database.

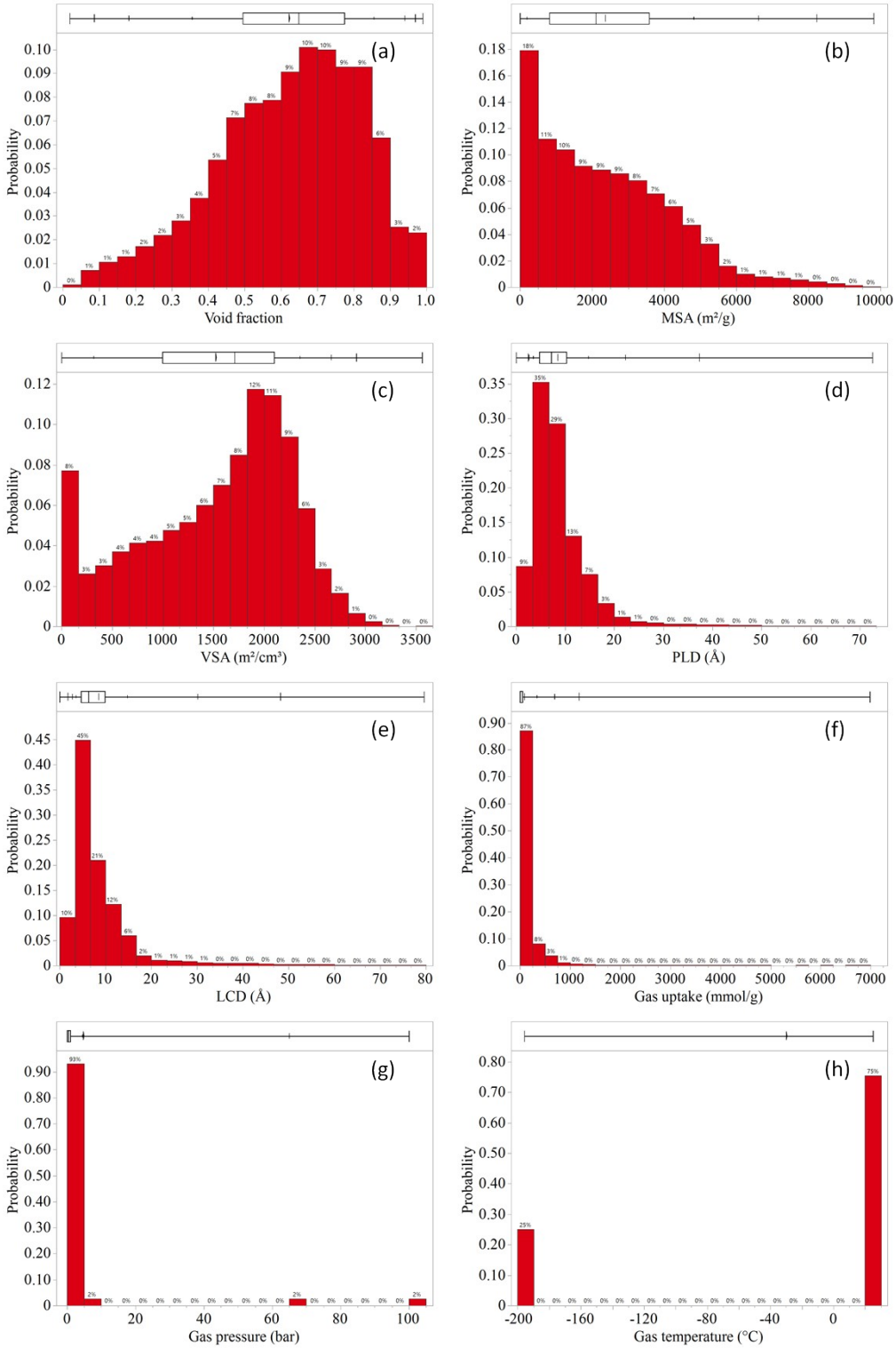

Figure S10 Histograms representing the probability distribution of key MOF descriptors in the screened database. Boxplots above each histogram provide additional statistical insights. The figures include: (a) Void Fraction, (b) MASA (m^2/g): Accessible surface area distribution, (c) VASA (m^2/cm^3): Surface area per unit volume, (d) PLD (Å): Pore-limiting diameter distribution, (e) LCD (Å): Largest cavity diameter, (f) Gas Uptake (mmol/g), (g) Gas Pressure (bar), and (h) Gas Temperature (°C).

Table S1 Hyperparameters and their tuning methods for machine learning algorithms.

Algorithm	Hyperparameters Tuned	Tuning Methods
Artificial Neural Network (ANN)	- Number of hidden layers - Number of neurons per layer (1-256) - Activation functions - Learning rate - Batch size - Epochs - Dropout rate - Optimizer	- Random Search - Grid Search - Bayesian search - Optuna - Ensemble Modeling
Random Forest	- Number of estimators (trees) (10-200) - Maximum depth (0-30) - Minimum samples split (2-10) - Minimum samples leaf (1-4) - Max features (sqrt-log2)	- Grid Search
k-Nearest Neighbors (k-NN)	- Number of neighbors (k) (1-30) - Distance metric - Weights	- Grid Search
Decision Tree	- Maximum depth (0-25) - Minimum samples split (5-300) - Minimum samples leaf (1-400) - Max features (sqrt-log2)	- Grid Search
LightGBM (LGBM)	- n estimators (50-400) -Learning rate (0.01-0.2) - number of leaves (20-50) - Alpha (0-1) - lambada (0-1) - Min_child_samples (5-20)	- Random Search - Grid Search
XGBoost (XGB)	- n estimators (50-400) - Learning rate (0.01-0.2) - Gamma (0-0.3)	- Random Search - Grid Search
Extra Trees (ET)	- Number of estimators (100-200) - Max depth (0-30) - Min samples split (2-7) - Min samples leaf (1-4) - Max feature (sqrt, log2, none)	- Grid Search
AdaBoost (AB)	- Learning rate (0.01-0.5) - Number of estimators (100-200) - Loss function (linear, square, exponential) - Estimator max depth - Estimator min depth	- Random Search
Gradient Boosting (GB)	- Estimators'(100-400) - Learning rate (0.05-0.2) - Max depth (3-9) - Min samples split (2-10) - Min samples leaf' (2- 6) - Subsample ratio (0.8-1.0) - Max features (sqrt, log2, None)	- Grid Search

Multiple Linear Regression (MLR)	- Regularization strength	- Grid Search
Ridge Regression	- Alpha (-6-100) - Solver ('auto', 'svd', 'cholesky', 'lsqr', 'sparse_cg', 'sag', 'saga')	- Random Search
Lasso Regression	- Alpha (-6-100)	- Grid Search

Table S2 Performance metrics of machine learning algorithms for water adsorption prediction.

	KH			W 0.1			W 0.2			W 0.3			W 0.4			W 0.5			W 0.6			W 0.7			W 0.8			W 0.9		
RF																														
Metric	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**
Train	0.996	8.430	25.030	0.993	0.001	0.003	0.996	0.001	0.003	0.998	0.003	0.013	0.998	0.003	0.010	0.998	0.003	0.010	0.998	0.005	0.015	0.999	0.006	0.017	0.999	0.027	0.072	0.998	0.008	0.023
Test	0.965	103.83	719.31	0.948	0.080	0.999	0.969	0.063	0.674	0.980	0.138	1.273	0.968	0.283	2.617	0.958	0.285	2.564	0.963	0.417	4.374	0.978	0.477	6.094	0.978	0.660	9.585	0.980	0.245	1.825
KNN																														
Train	0.993	8.430	25.030	0.993	0.001	0.003	0.996	0.001	0.003	0.997	0.003	0.013	0.994	0.003	0.010	0.992	0.003	0.010	0.993	0.005	0.015	0.995	0.006	0.017	0.996	0.027	0.072	0.996	0.008	0.023
Test	0.946	148.85	1031.1	0.940	0.091	1.137	0.962	0.074	0.785	0.969	0.207	1.913	0.945	0.441	4.076	0.931	0.471	4.235	0.937	0.707	7.417	0.952	0.861	10.999	0.955	1.132	16.448	0.970	0.324	2.418
LGBM																														
Train	0.996	8.430	25.030	0.995	0.001	0.003	0.997	0.001	0.003	0.999	0.003	0.013	0.999	0.003	0.010	0.999	0.003	0.010	0.999	0.005	0.015	0.999	0.006	0.017	0.999	0.027	0.072	0.998	0.008	0.023
Test	0.971	91.053	630.77	0.948	0.080	1.002	0.974	0.055	0.589	0.985	0.135	1.249	0.979	0.206	1.910	0.973	0.186	1.677	0.975	0.283	2.971	0.978	0.472	6.030	0.978	0.647	9.401	0.982	0.225	1.679
XGB																														
Train	0.996	8.430	25.030	0.993	0.001	0.003	0.997	0.001	0.003	0.999	0.003	0.013	0.999	0.003	0.010	0.999	0.003	0.010	0.999	0.005	0.015	0.999	0.006	0.017	0.999	0.027	0.072	0.998	0.008	0.023
Test	0.953	131.53	911.23	0.905	0.136	1.694	0.969	0.064	0.680	0.973	0.185	1.715	0.952	0.394	3.641	0.946	0.368	3.308	0.958	0.473	4.963	0.976	0.506	6.468	0.976	0.702	10.199	0.976	0.277	2.068
DT																														
Train	0.987	9.884	29.348	0.969	0.002	0.006	0.989	0.001	0.003	0.993	0.003	0.013	0.984	0.004	0.013	0.985	0.003	0.010	0.983	0.005	0.015	0.991	0.006	0.017	0.990	0.028	0.074	0.993	0.008	0.023
Test	0.930	186.67	1293.1	0.868	0.185	2.300	0.929	0.126	1.337	0.945	0.370	3.422	0.902	0.730	6.756	0.883	0.795	7.154	0.889	1.251	13.135	0.932	1.157	14.776	0.933	1.581	22.975	0.954	0.448	3.341
ET																														
Train	0.996	8.430	25.030	0.994	0.001	0.003	0.997	0.001	0.003	0.999	0.003	0.013	0.999	0.003	0.010	0.999	0.003	0.010	0.999	0.005	0.015	0.999	0.006	0.017	0.999	0.027	0.072	0.998	0.008	0.023
Test	0.974	82.881	574.16	0.965	0.059	0.735	0.979	0.048	0.509	0.987	0.135	1.249	0.980	0.202	1.864	0.979	0.145	1.307	0.980	0.230	2.415	0.986	0.356	4.548	0.987	0.867	6.790	0.985	0.199	1.487
AB																														
Train	0.985	10.418	30.933	0.978	0.002	0.005	0.980	0.001	0.004	0.985	0.003	0.013	0.971	0.006	0.019	0.966	0.005	0.017	0.976	0.006	0.018	0.980	0.009	0.026	0.977	0.045	0.121	0.990	0.008	0.023
Test	0.978	73.797	511.23	0.968	0.055	0.682	0.975	0.054	0.568	0.983	0.135	1.249	0.968	0.284	2.628	0.962	0.257	2.312	0.972	0.313	3.288	0.974	0.533	6.803	0.972	0.783	11.382	0.987	0.185	1.383
GB																														
Train	0.995	8.430	25.030	0.993	0.001	0.003	0.996	0.001	0.003	0.998	0.003	0.013	0.998	0.003	0.010	0.998	0.003	0.010	0.997	0.005	0.015	0.998	0.006	0.017	0.998	0.027	0.072	0.997	0.008	0.023
Test	0.975	81.116	561.93	0.962	0.063	0.785	0.978	0.050	0.531	0.989	0.135	1.249	0.982	0.190	1.762	0.980	0.139	1.254	0.982	0.225	2.362	0.981	0.426	5.440	0.982	0.582	8.462	0.984	0.209	1.560
MLR																														
Train	0.111	379.05	1125.4	0.122	0.044	0.133	0.176	0.042	0.125	0.171	0.124	0.539	0.197	0.122	0.407	0.203	0.120	0.399	0.208	0.198	0.594	0.161	0.255	0.721	0.160	1.148	3.060	0.179	0.332	0.956
Test	0.108	2097.2	14528.	0.120	1.157	14.419	0.175	1.294	13.738	0.168	5.616	51.958	0.195	5.503	50.912	0.203	5.423	48.804	0.208	8.910	93.539	0.169	12.405	158.494	0.169	17.208	250.083	0.183	6.659	49.670
Ridge																														
Train	0.278	308.43	915.77	0.363	0.032	0.097	0.455	0.028	0.083	0.416	0.088	0.380	0.468	0.081	0.271	0.466	0.080	0.267	0.465	0.134	0.401	0.380	0.189	0.535	0.376	0.856	2.284	0.547	0.185	0.533
Test	0.259	1747.3	12104.	0.334	0.879	10.950	0.432	0.896	9.512	0.389	4.127	38.185	0.439	3.851	35.630	0.437	3.831	34.483	0.437	6.335	66.507	0.364	9.529	121.751	0.360	13.282	193.035	0.533	3.836	28.615
Lasso																														
Train	0.272	310.88	923.05	0.355	0.033	0.098	0.451	0.028	0.084	0.410	0.089	0.384	0.461	0.082	0.275	0.456	0.082	0.272	0.455	0.136	0.409	0.373	0.191	0.541	0.368	0.866	2.310	0.519	0.196	0.564
Test	0.253	1760.6	12196	0.325	0.891	11.098	0.430	0.899	9.542	0.388	4.133	38.240	0.433	3.895	36.037	0.428	3.890	35.006	0.428	6.430	67.504	0.358	9.624	122.956	0.354	13.417	194.991	0.506	4.058	30.268
ANN																														
Train	0.978	13.551	40.234	0.978	0.002	0.005	0.991	0.001	0.003	0.989	0.003	0.013	0.983	0.004	0.013	0.984	0.003	0.010	0.983	0.005	0.015	0.968	0.013	0.036	0.968	0.057	0.152	0.987	0.009	0.027
Test	0.976	78.990	547.20	0.979	0.040	0.495	0.990	0.031	0.329	0.989	0.135	1.249	0.987	0.158	1.463	0.989	0.136	1.224	0.987	0.225	2.362	0.962	0.703	8.981	0.961	0.993	14.435	0.983	0.221	1.645

Table S3 Leave-one-MOF-out (LOMO) cross-validation metrics for the stacked meta-model across all frameworks in the training set.

Target	Pooled R^2	Pooled MAE	Pooled RMSE
KH	0.93	98.51	402.2
W 0.1	0.95	0.043	0.456
W 0.2	0.94	0.101	0.819
W 0.3	0.96	0.255	2.043
W 0.4	0.95	0.257	3.745
W 0.5	0.98	0.192	2.443
W 0.6	0.98	0.225	2.362
W 0.7	0.97	0.871	5.178
W 0.8	0.96	0.865	7.655
W 0.9	0.97	1.544	3.656

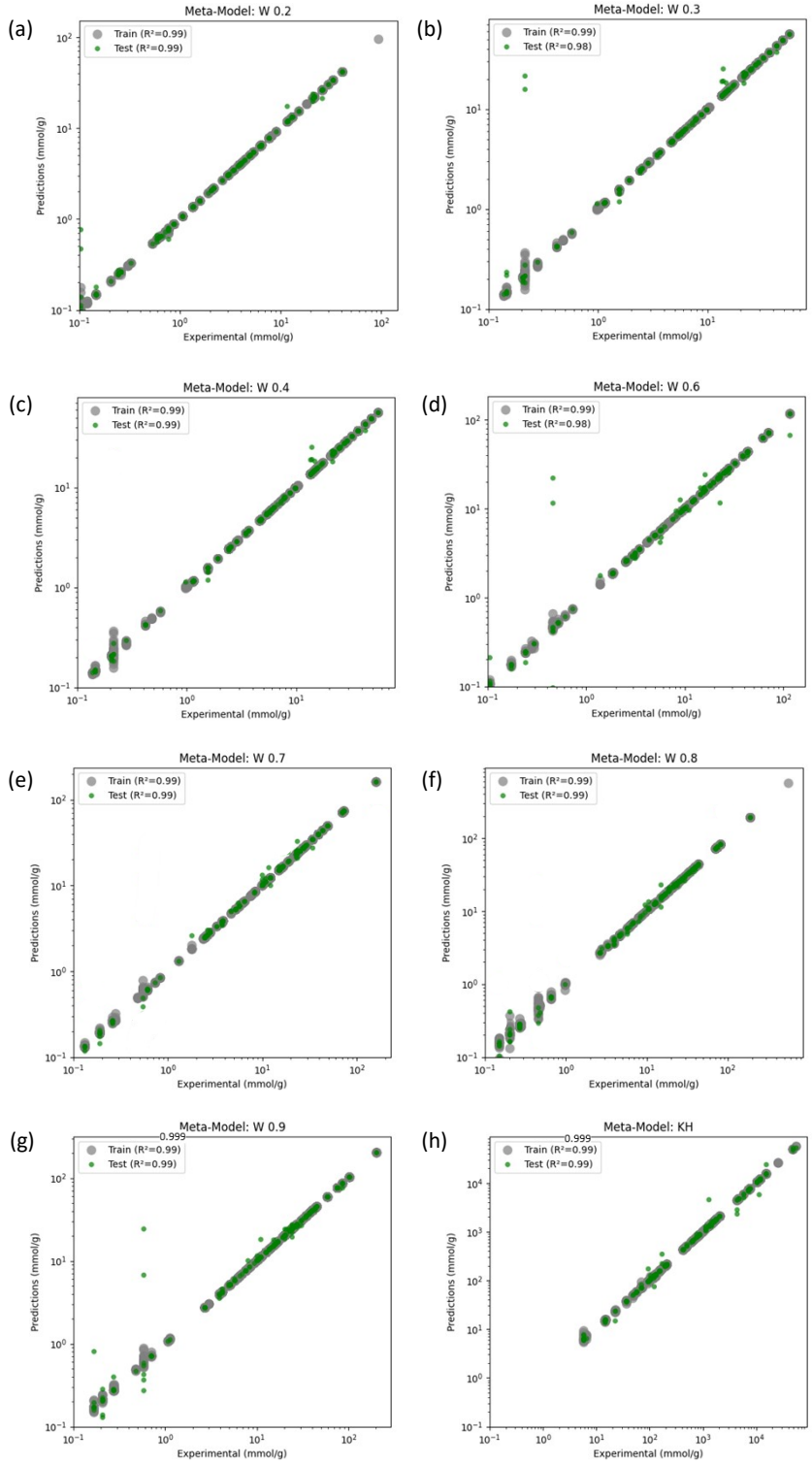

Figure S11 Parity plots for meta-model predictions comparing predicted values to experimental data on a log–log scale. Each subplot corresponds to a different target variable: (a) W 0.2, (b) W 0.3, (c) W 0.4, (d) W 0.6, (e) W 0.7, (f) W 0.8, (g) W 0.9, and (h) KH (Henry's Constant). Green and gray dots represent test and train predictions, respectively, with R^2 values annotated in the legends.

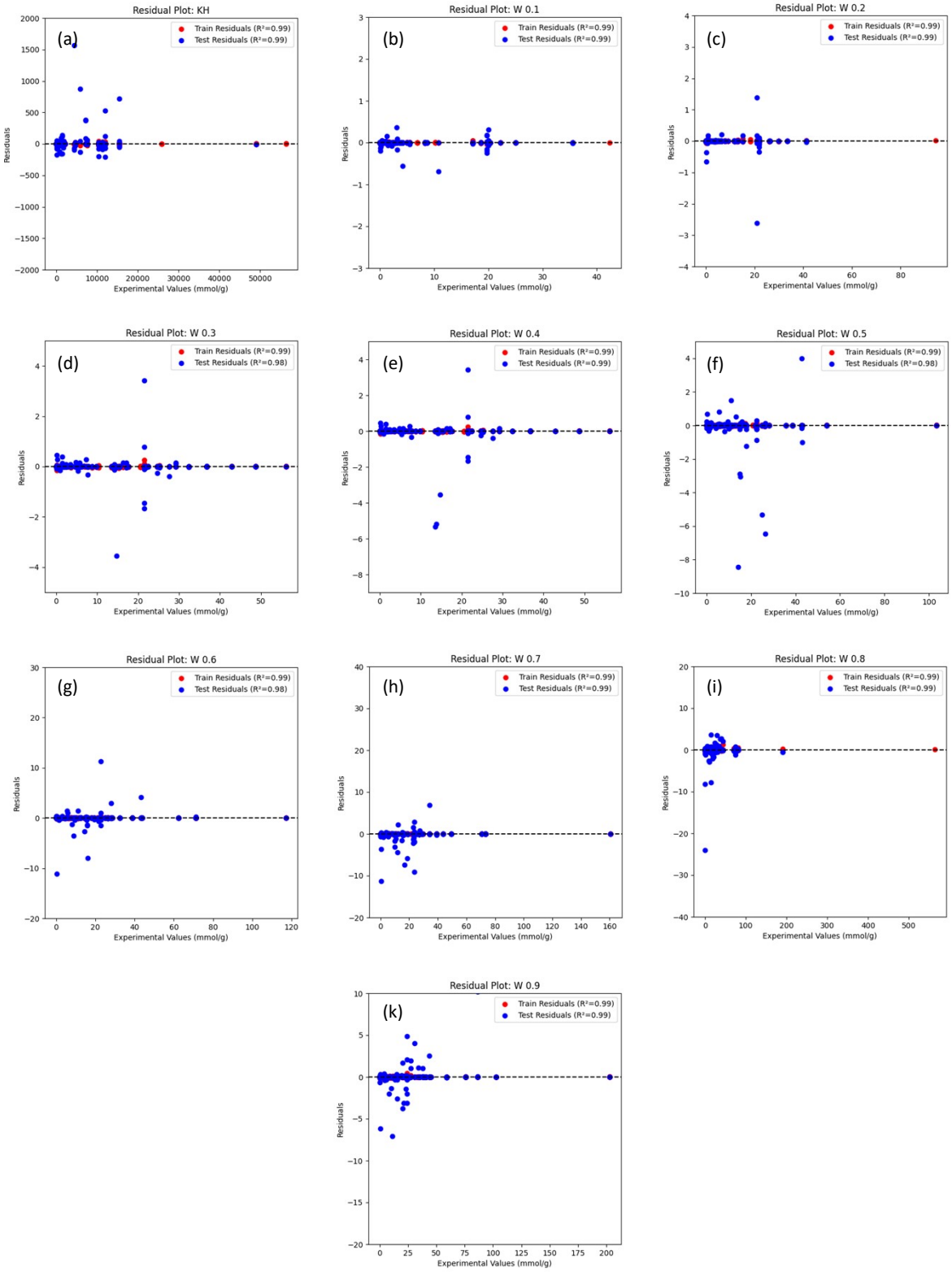

Figure S12 Residual plots for meta-model predictions comparing experimental values to residuals. Each subplot represents a specific target variable: (a) KH (Henry's Constant), (b) W 0.1, (c) W 0.2, (d) W 0.3, (e) W 0.4, (f) W 0.5, (g) W 0.6, (h) W 0.7, (i) W 0.8, and (k) W 0.9. Red points represent training residuals, and blue points represent test residuals. The dashed black line indicates zero residual error. High R^2 values for both training and testing indicate strong model accuracy with minimal bias.

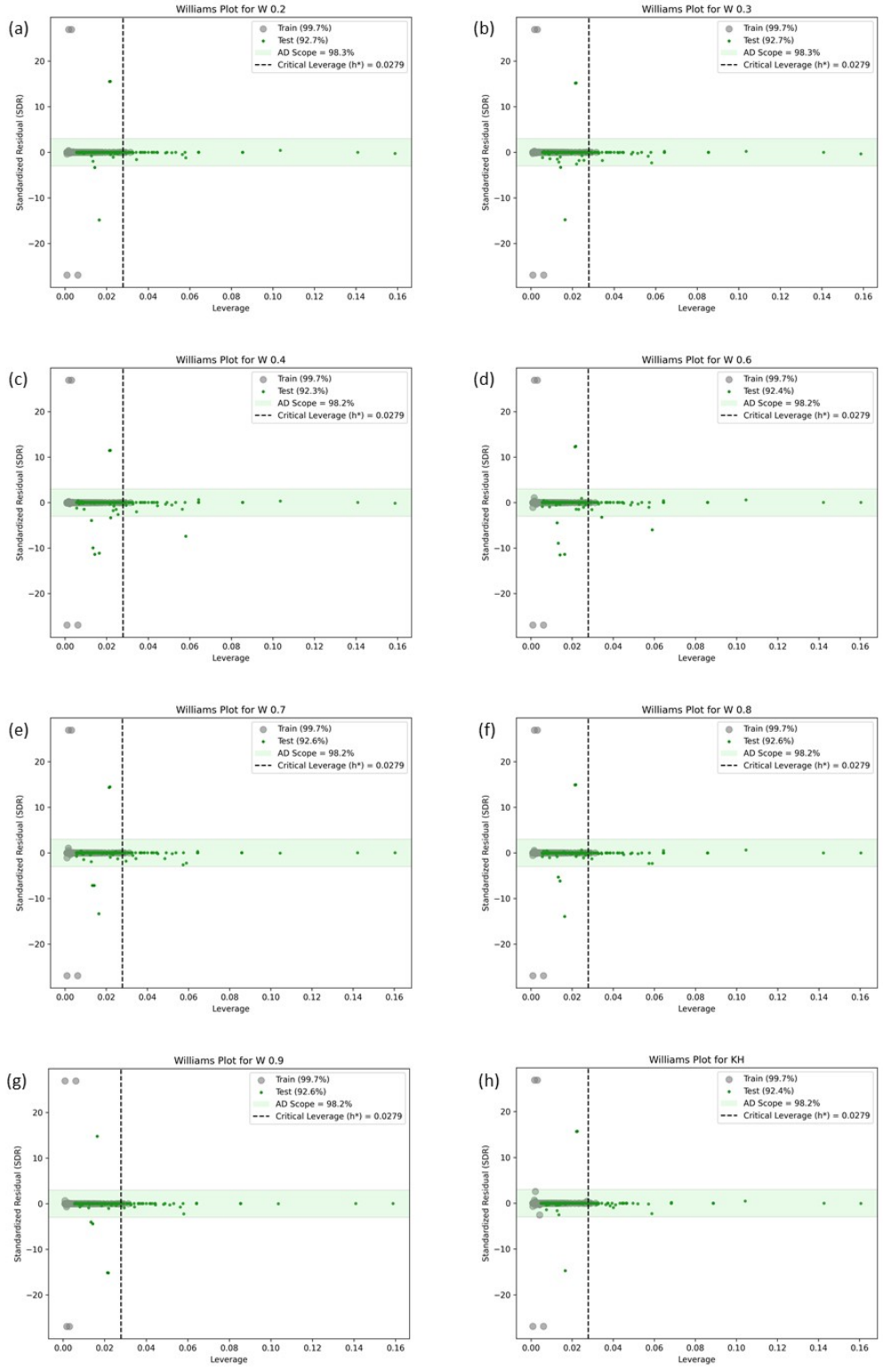

Figure S13 A series of Williams plots evaluating the applicability domain (AD) of the meta-model for various adsorption properties. Each plot assesses standardized residuals (SDR) versus leverage values, identifying response and structural outliers for KH (Henry's Constant) and W0.1 – W0.9.

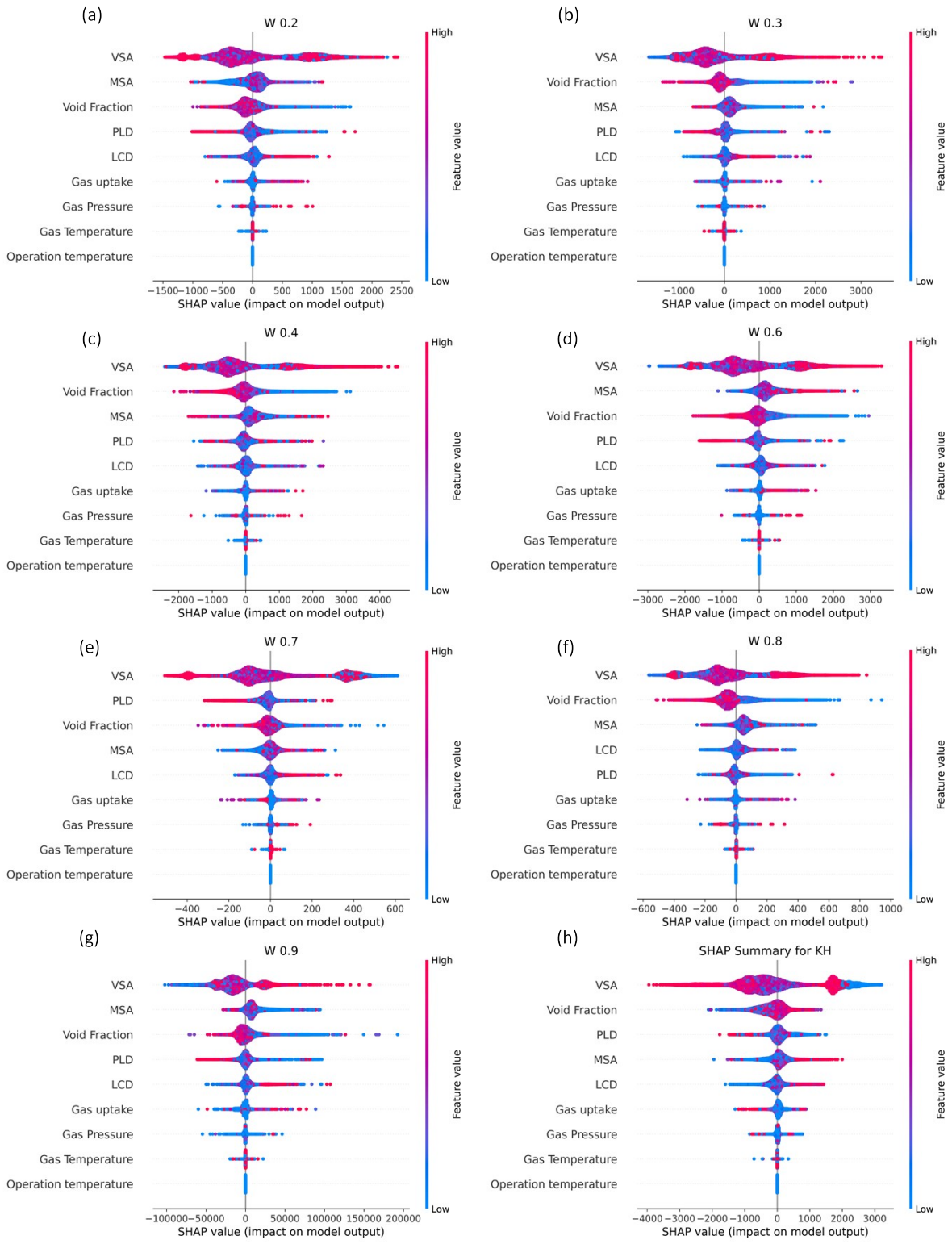

Figure S14 SHAP Analysis for Feature Importance in Meta-Model Predictions, illustrating the contribution of various structural and adsorption-related descriptors to the meta-model's predictions for different adsorption properties (KH and W0.1–W0.9). Each plot highlights the positive or negative influence of features such as LCD, Void Fraction, MASA, and gas properties on the adsorption performance of MOFs.

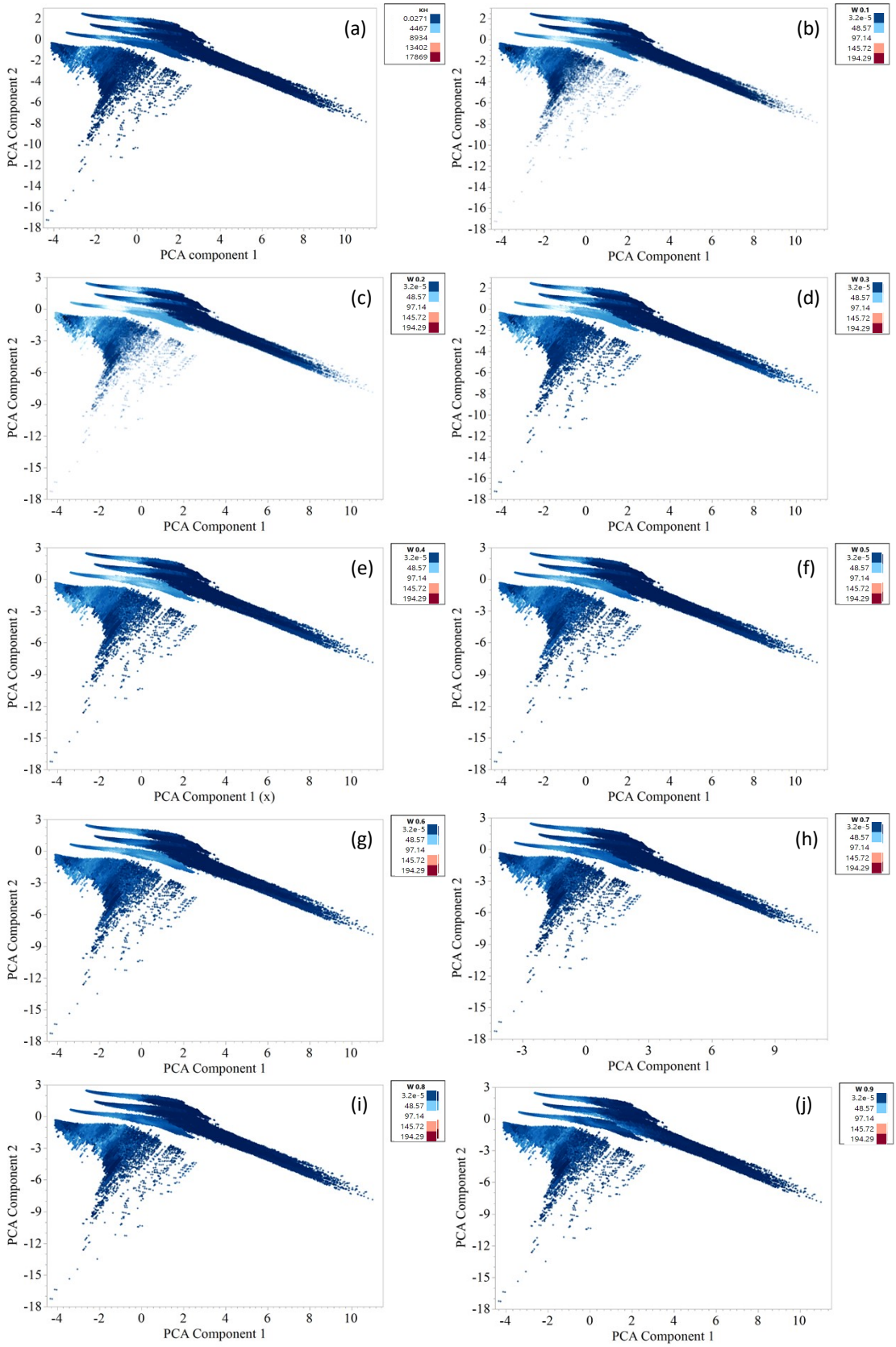

Figure S15 Bubble plots showing the principal component analysis (PCA) projections of MOF adsorption properties. Each plot represents a different adsorption metric (KH, W0.1–W0.9) with color-coded values, highlighting material clusters and key property distributions in PCA space.

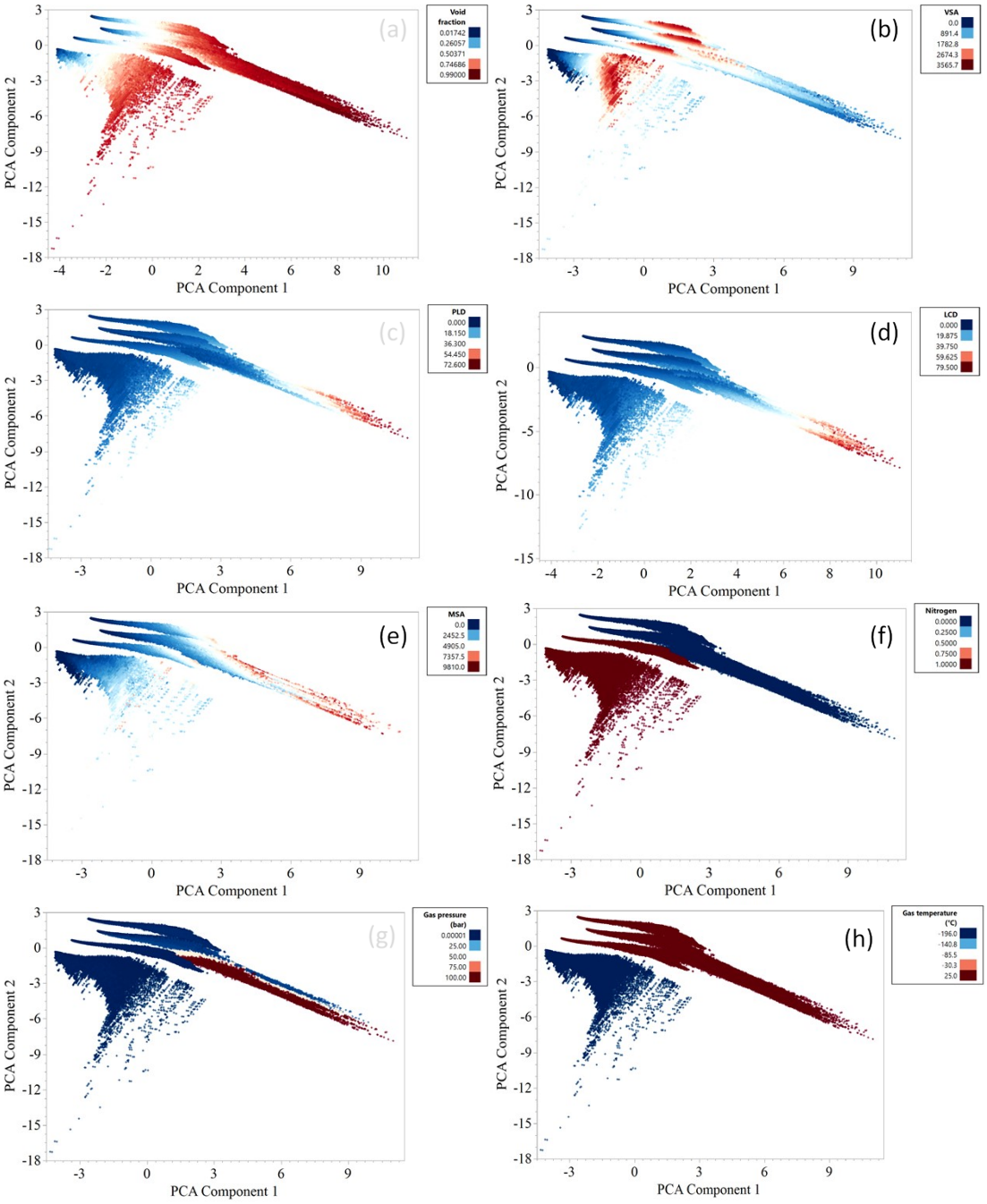

Figure S16 Bubble plots depicting PCA projections of MOF materials, color-coded by different structural descriptors.

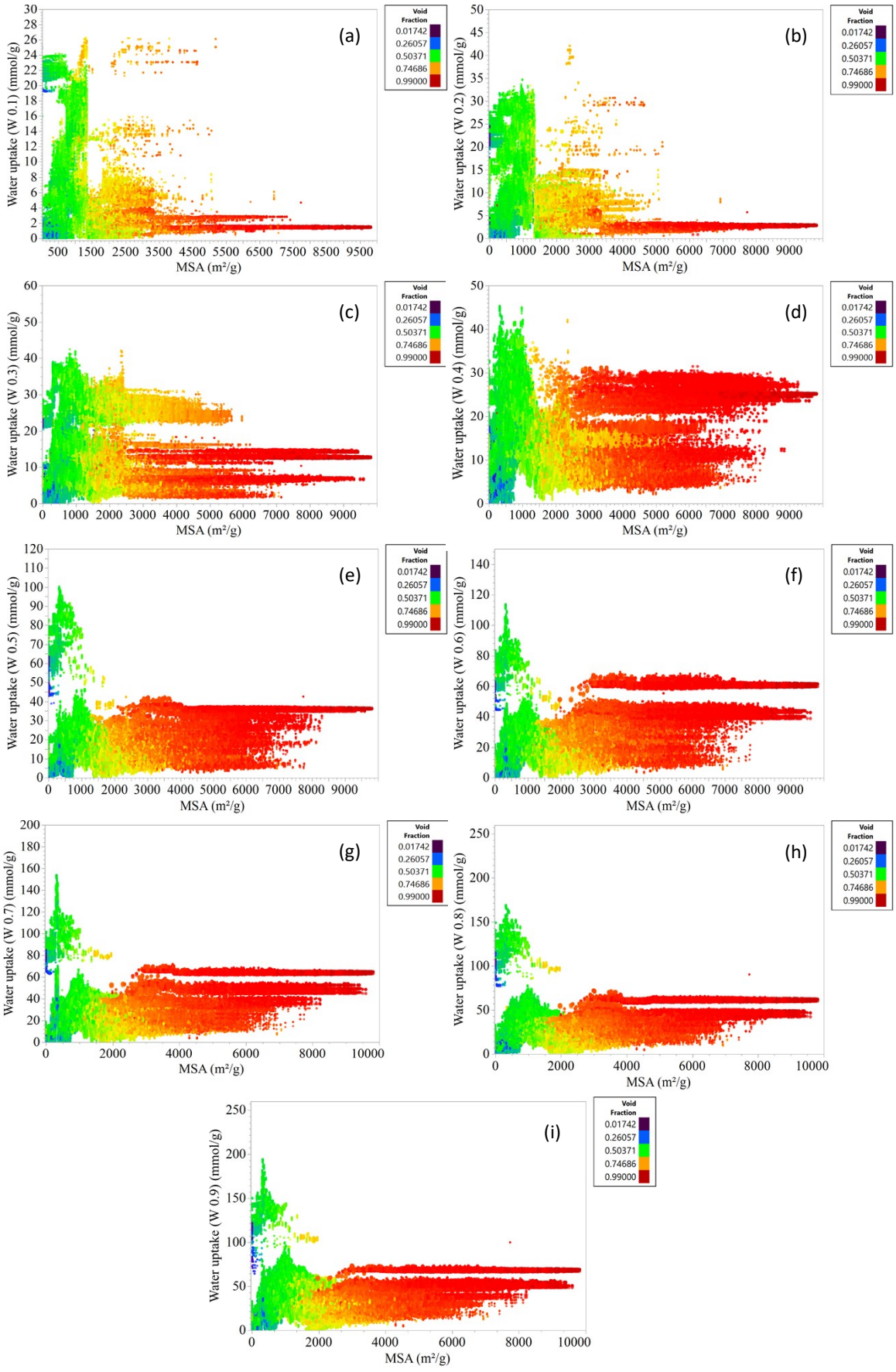

Figure S17 Bubble Plots of KH and W0.1–W0.9 by MSA (m²/g), Sized by PLD (Å).

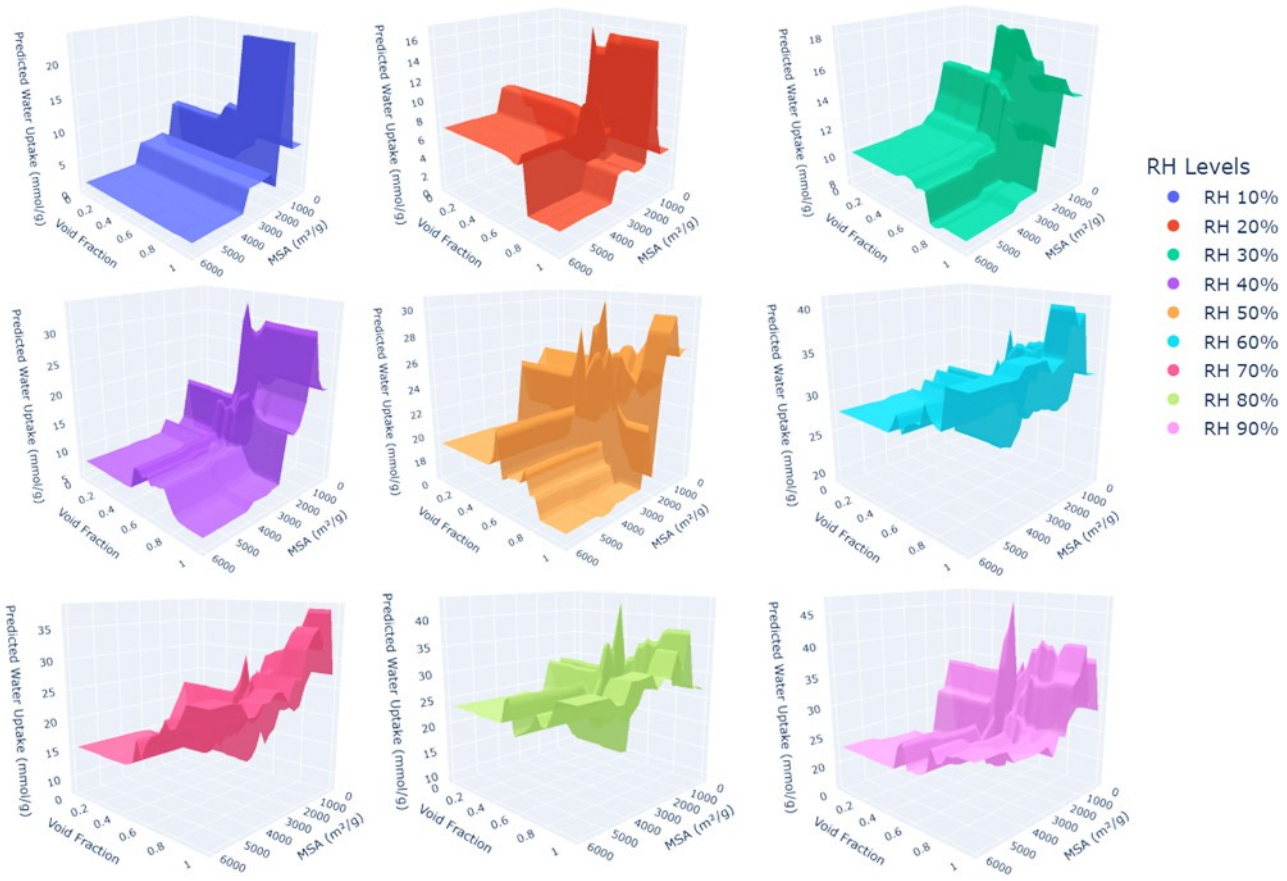

Figure S18 3D surface plots of predicted water uptake (mmol/g) versus surface area and Void Fraction across RH levels (10–90%).

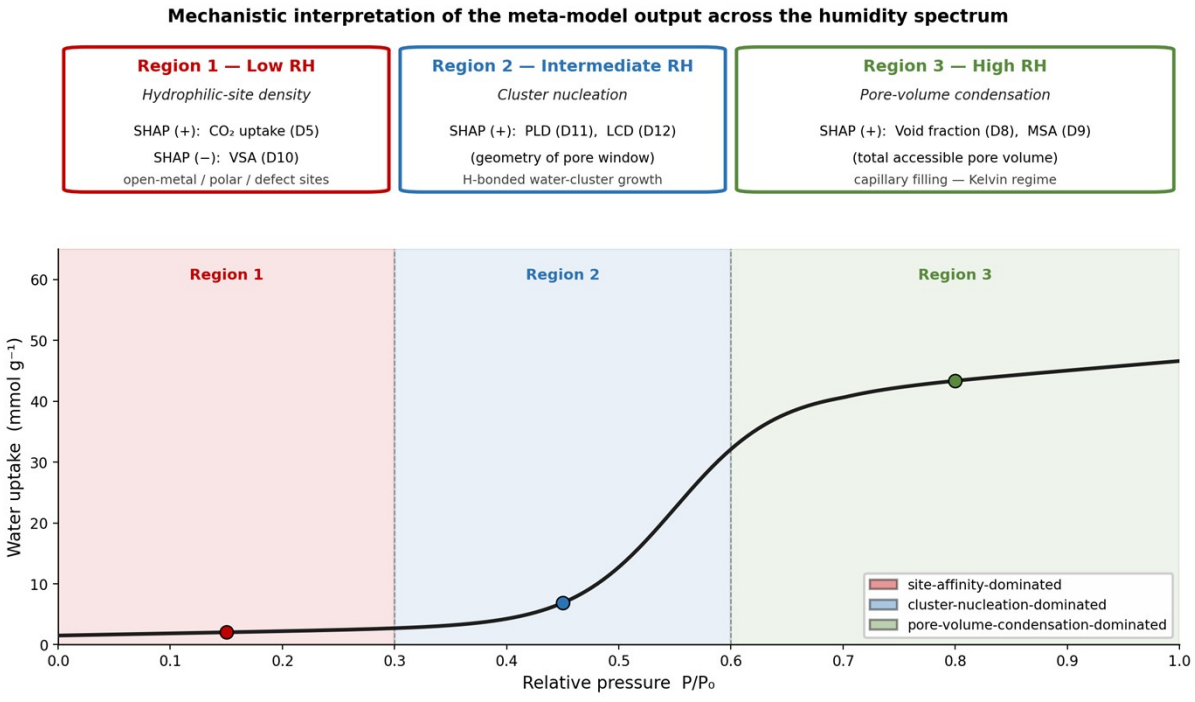

Figure S19 Schematic summarizing the humidity-resolved adsorption regimes inferred from the SHAP, PCC and 3D descriptor-map analyses: at low RH the uptake is dominated by hydrophilic-site density, captured by the model through the positive SHAP contribution of CO_2 uptake (D5) and the negative contribution of VSA (D10); at intermediate RH by hydrogen-bonded cluster nucleation, captured by the positive SHAP contributions of PLD (D11) and LCD (D12); and at high RH by pore-volume-driven capillary condensation, captured by the positive SHAP contributions of void fraction (D8) and MSA (D9).

Table S4 Chemistry-group test of the proxy-descriptor framework. The training MOFs were classified by dominant linker functional class using CSD-known linker chemistry and grouped into carboxylate-rich, hydroxyl-bearing, mixed-functionality, and aromatic-only / hydrophobic classes. For each class, the table reports median water uptake at W0.5 and W0.9, the interquartile range of W0.9, and the median Henry constant KH. The aromatic-only/hydrophobic class shows markedly lower KH and W0.9 values than the hydrophilic-linker classes, supporting the conclusion that the hybrid descriptor set captures the broad hydrophilic/hydrophobic distinction relevant to AWH. Within the hydrophilic classes, the medians are broadly comparable, while the carboxylate-rich class spans the widest uptake range, consistent with the broader tunability of carboxylate frameworks rather than with any single linker chemistry acting as the sole determinant of uptake.

Linker functional class	**Examples / typical groups**	**MOFs in class (of 54)**	**Median W 0.5 (mmol g^{-1})**	**Median W 0.9 (mmol g^{-1})**	**IQR W 0.9 (mmol g^{-1})**	**Median KH (mmol bar^{-1} g^{-1})**
Carboxylate-rich	BDC, BTC, BTBA, biphenyl-COOH at metal nodes	25	8.1	15.3	8.6 – 34.6	851
Hydroxyl-bearing	–OH, on linker, μ-OH bridges (Al, Zr, Cr nodes)	13	13.2	19.4	10.4 – 37.8	867
Mixed functionality	$–COO^-$ + $–NH_2$ / $–COO^-$ + pyridyl / charged frameworks	7	10.1	17.0	7.9 – 27.7	1333
Aromatic-only / hydrophobic	Predominantly aromatic linkers, no polar functional groups	9	0.4	0.71	0.27 – 1.13	50

Table S5 CO_2-uptake stratification as a direct failure-mode check on the proxy approach. The subset of training MOFs with measurable CO_2 adsorption data was ranked by median CO_2 uptake and grouped into three terciles. For each tercile, the table reports the CO_2 uptake range, the mean low-RH water uptake (W0.1–W0.3), the mean high-RH water uptake (W0.7–W0.9), and the median Henry constant KH. Low-RH water uptake increases from the lowest-CO_2 tercile to the middle tercile but saturates in the highest-CO_2 tercile because a small number of frameworks adsorb CO_2 strongly while remaining weak water adsorbents. These cases define an important boundary of the proxy approach and motivate the use of applicability-domain leverage as a practical reliability indicator for chemically atypical frameworks.

CO_2-uptake tercile	CO_2 uptake range (mmol g^{-1})	MOFs (of 27 with CO_2 data)	Mean W 0.1 – W 0.3 (mmol g^{-1})	Mean W 0.7 – W 0.9 (mmol g^{-1})	Median KH (mmol bar^{-1} g^{-1})
T1 (lowest CO_2)	0.04 – 1.13	9	3.6	16.6	195
T2 (mid CO_2)	1.30 – 2.76	9	10.9	37.1	1435
T3 (highest CO_2)	4.76 – 9.98	9	9.3	19.4	1747

The relationship is monotonic from T1 to T2 but saturates at T3 because of two MOFs (TASXIW_clean and GUXJUF_clean) that adsorb CO_2 strongly through CO_2-specific binding sites that exclude water; these are exactly the boundary cases the applicability-domain leverage flag ($h^* = 0.0279$) is designed to surface for experimental verification.

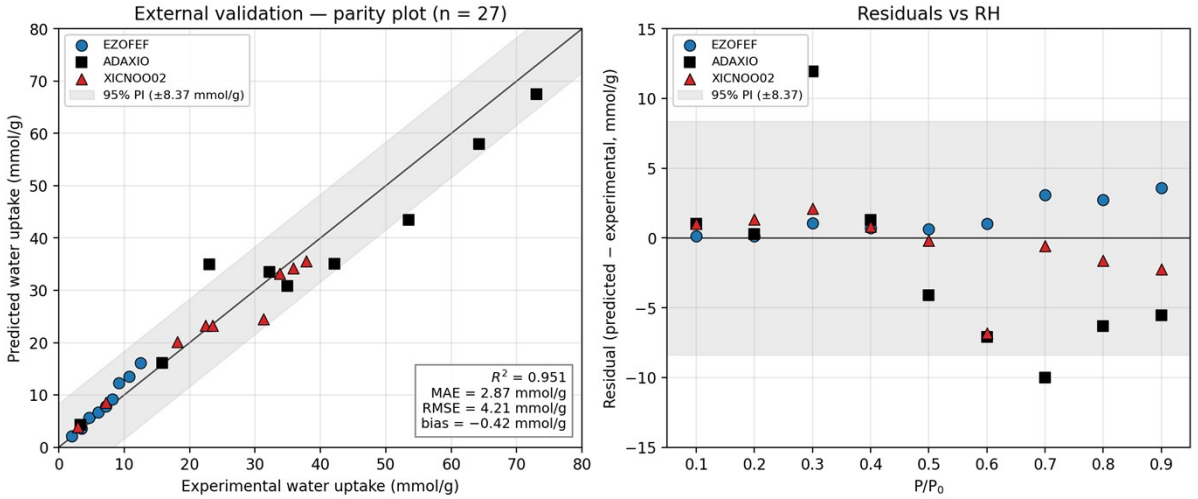

Figure S20 External validation of the meta-model against three independent water adsorption isotherms not included in the training set: EZOFEF (NIST/ARPA-E adsorption database), and ADAXIO and XICNOO02 (measured experimentally in this work at 25 °C). Each isotherm contributes nine paired predicted–measured points across $P/P_0 = 0.1–0.9$, giving $n = 27$ external test points. (Left) Parity plot of predicted versus experimental water uptake. The diagonal line indicates perfect agreement; the shaded band represents the empirical 95 % prediction interval (±8.37 mmol/g) constructed from the pooled external residuals. The inset reports the pooled metrics: $R^2 = 0.951$, MAE = 2.87 mmol/g, RMSE = 4.21 mmol/g, bias = −0.42 mmol/g. (Right) Residuals (predicted − experimental) as a function of relative pressure, with the same 95 % prediction interval shown as the shaded band. All 27 points lie within the prediction interval except for ADAXIO at $P/P_0 = 0.7$, where the model under-predicts the steep capillary-condensation step by ≈10 mmol/g; this corresponds to the model placing the condensation step approximately one RH bin earlier than experiment. The residuals show no systematic trend with RH at low to intermediate humidity but a mild negative bias at high RH ($P/P_0 \geq 0.6$), consistent with a small under-prediction of the saturation plateau in highly hydrophilic frameworks.

Table S6 Empirical 95 % prediction intervals for the meta-model. RMSE_LOMO values are taken from Table S3 (n = 64 LOMO folds); external-validation PIs are computed from the residuals of the three external MOFs (EZOFEF, ADAXIO, XICNOO02; n = 27). The pooled ±8.37 mmol/g entry is the recommended single-number uncertainty envelope for prospective predictions. Units: KH in mmol·bar^{-1}·g^{-1}; W targets in mmol/g.

Target	**RMSE_LOMO**	**±1.96·RMSE_LOMO**	**External 95 % PI**
KH	402.2	±788	-
W 0.1	0.456	±0.89	±1.00
W 0.2	0.819	±1.61	±1.27
W 0.3	2.043	±4.00	±11.79
W 0.4	3.745	±7.34	±0.60
W 0.5	2.443	±4.79	±4.96
W 0.6	2.362	±4.63	±9.02
W 0.7	5.178	±10.15	±13.22
W 0.8	7.655	±15.00	±8.90
W 0.9	3.656	±7.17	±9.05
Pooled W	-	-	±8.37

Table S7 Top-performing MOFs from the screened database.

MOF	Void Fraction	MSA (m²/g)	VSA (m²/cm³)	PLD (Å) (mmol/g)	LCD (Å)	KH (mmol/bar.g)	W 0.1 (mmol/g)	W 0.2 (mmol/g)	W 0.3 (mmol/g)	W 0.4 (mmol/g)	W 0.5 (mmol/g)	W 0.6 (mmol/g)	W 0.7 (mmol/g)	W 0.8 (mmol/g)	W 0.9 (mmol/g)
EBITAL_clean	0.456	314.347	476.219	2.51846	4.48474	10736.92757	22.28266	29.83381	37.79288	44.69993	99.09139	112.4981	152.5099	167.7861	193.0438
FOQTAH_clean	0.4578	356.701	474.826	2.70699	4.13249	10736.92757	21.94471	29.57654	36.99343	38.64457	94.6281	108.6319	145.2626	157.939	186.8238
NICCAE_clean	0.455	330.811	580.072	2.75227	4.49529	10736.92757	21.97171	29.56767	37.80853	43.3149	95.59219	106.1657	149.4739	165.7766	186.0983
TUKCUZ_clean	0.4628	359.371	498.673	2.40262	4.88707	10585.8366	21.98761	29.56266	37.69531	41.50761	93.23718	101.8115	135.8228	148.7735	184.3772
CANZOH_clean	0.4554	307.142	547.931	2.47537	4.29669	10736.92757	22.26465	29.87857	37.19184	37.56142	90.63996	102.9606	143.3297	156.5883	180.6661

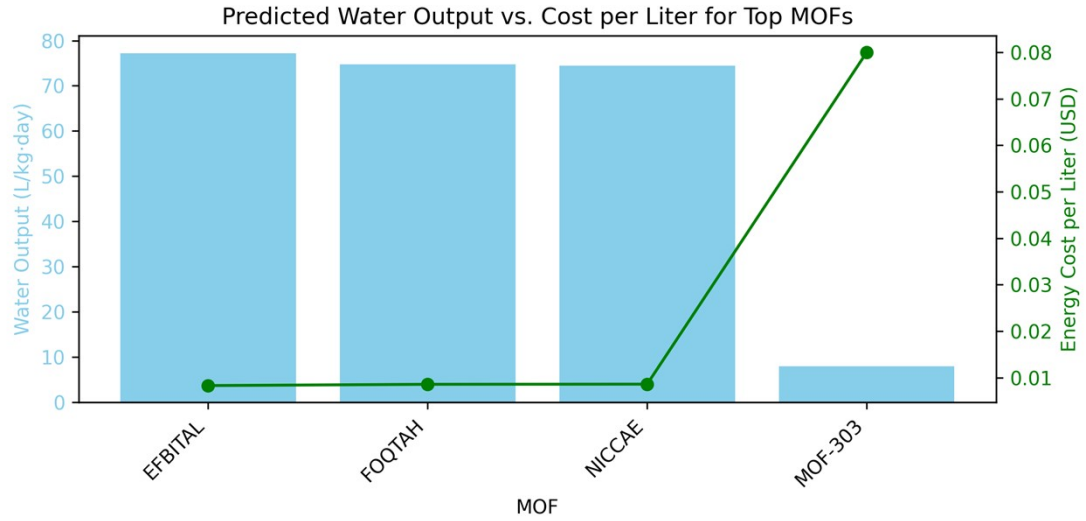

Figure S21 Comparison of daily water output and cost per liter for selected high-performing MOFs and the MOF-303 benchmark used in a commercial system.

Table S8 Benchmark AWH MOFs and retained their initial uptake within experimental uncertainty.

Material	Uptake at 30 % RH (mmol/g)	Peak uptake (mmol/g)	Working RH window	Regen. T (°C)	Cycling evidence
MOF-303	23	27 (40 % RH)	10–35 % RH	65–85	> 100 cycles
MOF-801	27	35 (50 % RH)	5–35 % RH	65–85	> 80 cycles
Co_2Cl_2(BTDD)	55	55 (40 % RH)	20–35 % RH	65–85	> 10 cycles
CAU-10-H	27	31 (50 % RH)	15–40 % RH	65–95	> 700 cycles
EZOFEF	4.6	12.5 (90 % RH)	10–90 % RH	< 85	10 cycles, stable (this work)
ADAXIO	23	73.1 (90 % RH)	20–90 % RH	< 85	10 cycles, stable (this work)
XICNOO02	18	37.9 (90 % RH)	20–90 % RH	< 85	10 cycles, stable (this work)

Table S9 Methodological benchmarking of representative machine-learning workflows for MOF property prediction and screening relevant to adsorption and materials discovery.

Representative study	Primary target / application	Data foundation	Representation / descriptor strategy	Workflow type	Validation / reliability treatment	Practical contribution	Position relative to the present work
Li et al., Nanomaterials 2022 (*16*)	AWH-oriented computational screening of water capture from air	6,013 CoRE-MOFs + 137,953 hypothetical MOFs from simulation	Six hand-crafted descriptors; (Qst) highlighted as key descriptor	Descriptor-based supervised screening	Grid search + five-fold CV; (R^2) up to 0.97 on CoRE-MOFs and 0.86 transfer to hMOFs	Strong example of simulation-first AWH screening	Closest prior AWH benchmark, but more simulation-driven and less experimentally anchored to measured water isotherms than the present work

Rosen et al., Matter 2021 (QMOF) (*17*)	Quantum-chemical property prediction and accelerated electronic-material discovery in MOFs	>14,000 experimentally synthesized MOFs; >170 years of compute time behind the database	Structure-derived ML surrogates trained on DFT-quality labels	First-principles + ML screening	Benchmarked against computed quantum labels	Strong example of physics/DFT-integrated ML	Much stronger quantum fidelity, but focused on electronic properties rather than water adsorption / AWH
Nandy et al., Sci Data 2022 (MOFSimplify) (*18*)	MOF stability prediction	>2,000 solvent-removal stability records + 3,000 thermal decomposition temperatures from literature extraction	Graph- and pore-geometry-based features / RAC-style fingerprints	Experimental-data curation + ML + web interface	Quantified uncertainty; community-feedback loop for active learning	Strong example of uncertainty-aware and deployment-oriented MOF ML	Highly complementary to the present work as a secondary stability filter after adsorption screening
Kang et al., Nat Mach Intell 2023 (MOFTransformer) (*19*)	Universal transfer learning across MOF properties	Pretrained on 1 million hypothetical MOFs; fine-tuned on datasets of 5,000–20,000 MOFs	Multimodal transformers with atom-based graph + energy-grid embeddings	Foundation / transfer-learning workflow	Multi-property benchmarking; attention-based interpretation	Strong example of large-scale representation learning	Broader and more universal than the present work, but less specifically tied to experimentally measured water-isotherm behavior
Wang et al., Nat Commun 2024 (Uni-MOF) (*20*)	Unified gas adsorption prediction across gases and operating conditions	>631,000 MOF/COF structures; large cross-system adsorption datasets	Pure 3D structural representation from CIF + gas / T / P conditions	Self-supervised representation learning for adsorption	Broad-condition benchmarking; correspondence with experiments reported	Strong example of a CIF-first universal gas-adsorption estimator	More general for gas adsorption than the present work, but not focused on experimental water-isotherm reconstruction for AWH

Khan & Moosavi, Nat Commun 2025 (*21*)	Connecting newly synthesized MOFs to applications	Pretraining on structure databases; deployment from PXRD + synthesis precursors available immediately after synthesis	Multimodal PXRD-spectrum + precursor-string model	Synthesis-to-application recommendation	Robustness checks against experimental imperfections; "time-travel" test correctly identified 16/18 carbon-capture candidates	Strong example of low-friction post-synthesis deployment	Conceptually close in practical deployment emphasis, but aimed at application recommendation rather than water-isotherm prediction
Colón et al., Chem Sci 2024 (*22*)	Reducing simulation cost for adsorption modeling in porous materials	Active-learning campaign spanning 1,800 MOFs	Alchemical-molecule surrogates + MLP adsorption model	Active-learning adsorption workflow	Reduced training data by 57.5% while retaining predictive accuracy	Strong example of data-efficient iterative learning	Best viewed as a natural future extension for the present workflow rather than a direct benchmark on the same target
Qiu et al., Mater. Today (2025) (*24*)(*24*)	Recent advances and methodological trends in ML-assisted MOF discovery, screening, and design (perspective / review)	Survey of representative simulation- and experiment-derived MOF datasets used in recent ML studies (CoRE-MOF, hypothetical MOFs, QMOF, Cambridge Structural Database extracts)	Critical synthesis of descriptor families used across recent MOF-ML work (geometric, energetic, graph-based, transformer embeddings)	Methodological perspective / review	Discusses validation strategies adopted across recent studies including cross-validation, simulation-derived benchmarking, and the emerging role of uncertainty quantification	Strong overview of the field's methodological trajectory; identifies open challenges including data scarcity for chemistry-specific properties and the need for experimentally anchored validation	Provides the broader methodological context against which the present work positions itself; the present workflow corresponds to the experimentally anchored, uncertainty-aware, deployment-oriented branch identified in this review as a needed direction
Qiu et al., Chem. Sci. (2025) (*23*)(*24*)	ML-integrated high-throughput screening of porous materials for gas separation / adsorption-driven applications	Combined simulation- and structure-derived dataset of porous materials with task-specific adsorption / separation labels	ML model trained on descriptor combinations integrating geometric, energetic, and chemical features for gas-adsorption prediction and ranking	ML + molecular-simulation-integrated screening workflow	Cross-validation against simulation-derived ground truth and reported correspondence with literature experimental data where available	Demonstrates the productivity of integrating ML with molecular simulations and high-throughput screening for accelerating discovery in gas-separation / adsorption applications	Methodologically aligned with the present work in coupling ML with simulation-derived descriptors at scale, but targets gas-separation / adsorption rather than experimentally anchored water-isotherm prediction; the present workflow contributes the AWH-specific, experimentally

							validated counterpart in the same broader ML-assisted screening paradigm
Present work	Water-isotherm prediction and AWH-focused MOF screening	88 experimental water isotherms across 64 MOFs, complemented with three gases sorption descriptors; deployed on a 554,000-record screening database spanning >54,000 MOFs	Hybrid structural + gas-sorption descriptor workflow	Stacked ensemble forward-screening model	LOMO, 3-MOF external validation, applicability-domain leverage, and empirical residual-based uncertainty interval	Strong example of an experimentally anchored, externally validated hybrid AWH workflow under data scarcity	Distinguishing niche: narrower than foundation models, but stronger on experimental water relevance, interpretability, and practical AWH screening

References

1. A. Mathai, S. Provost, H. Haubold, Chapter 9: Principal Component Analysis. *Multivariate Statistical Analysis in the Real and Complex Domains*, 597–639 (2022).
2. L. van der Maaten, G. Hinton, Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).
3. VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* **58**, 240–242 (1895).
4. L. Breiman, Random forests. *Mach. Learn.* **45**, 5–32 (2001).
5. T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning. doi: 10.1007/978-0-387-84858-7 (2009).
6. D. H. Wolpert, Stacked generalization. *Neural Networks* **5**, 241–259 (1992).
7. Z.-Hua. Zhou, Ensemble methods : foundations and algorithms. 222 (2012).
8. S. Džeroski, B. Ženko, Is combining classifiers with stacking better than selecting the best one? *Mach. Learn.* **54**, 255–273 (2004).
9. T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **13-17-August-2016**, 785–794 (2016).
10. D. Y. Y. Sim, Z. Wei, XGBoost Regression Algorithms for Efficient Predictions on Inventory Sales and Management. 66–71 (2024).
11. S. p Shaji, R. R, J. Varghese, L. Sathyan, D. J, Optimizing Hyperparameters: Techniques for Improving Machine Learning Models. *International Research Journal on Advanced Engineering and Management (IRJAEM)* **2**, 3782–3787 (2024).
12. M. Leblanc, R. Tibshirani, Combining Estimates in Regression and Classification. *J. Am. Stat. Assoc.* **91**, 1641–1650 (1996).
13. L. Breiman, Bagging predictors. *Mach. Learn.* **24**, 123–140 (1996).
14. C. Sutton, M. Boley, L. M. Ghiringhelli, M. Rupp, J. Vreeken, M. Scheffler, Identifying domains of applicability of machine learning models for materials science. *Nature Communications 2020 11:1* **11**, 4428- (2020).
15. S. M. Lundberg, S. I. Lee, A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017-December**, 4766–4775 (2017).
16. L. Li, Z. Shi, H. Liang, J. Liu, Z. Qiao, Machine Learning-Assisted Computational Screening of Metal-Organic Frameworks for Atmospheric Water Harvesting. *Nanomaterials* **12**, 159 (2022).
17. A. S. Rosen, S. M. Iyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J. M. Notestein, R. Q. Snurr, Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter* **4**, 1578–1597 (2021).

18. A. Nandy, G. Terrones, N. Arunachalam, C. Duan, D. W. Kastner, H. J. Kulik, MOFSimplify, machine learning models with extracted stability data of three thousand metal–organic frameworks. *Scientific Data 2022 9:1* **9**, 74- (2022).

19. Y. Kang, H. Park, B. Smit, J. Kim, A multi-modal pre-training transformer for universal transfer learning in metal–organic frameworks. *Nature Machine Intelligence 2023 5:3* **5**, 309–318 (2023).

20. J. Wang, J. Liu, H. Wang, M. Zhou, G. Ke, L. Zhang, J. Wu, Z. Gao, D. Lu, A comprehensive transformer-based approach for high-accuracy gas adsorption predictions in metal-organic frameworks. *Nature Communications 2024 15:1* **15**, 1904- (2024).

21. S. T. Khan, S. M. Moosavi, Connecting metal-organic framework synthesis to applications using multimodal machine learning. *Nature Communications 2025 16:1* **16**, 5642- (2025).

22. E. Osaro, F. Fajardo-Rojas, G. M. Cooper, D. Gomez-Gualdron, Y. J. Colon, Active learning of alchemical adsorption simulations; towards a universal adsorption model. *Chem. Sci.* **15**, 17671–17684 (2024).