Electronic Supplementary Material (ESI) for Journal of Materials Chemistry A.
This journal is © The Royal Society of Chemistry 2023

Supporting Information

for

Advancing CH_4/H_2 Separation with Covalent Organic Frameworks by Combining Molecular Simulations and Machine Learning

Gokhan Onder Aksu, Seda Keskin*

Department of Chemical and Biological Engineering, Koc University, Rumelifeneri Yolu, Sariyer, 34450, Istanbul, Turkey

Submitted to *Journal of Materials Chemistry A*

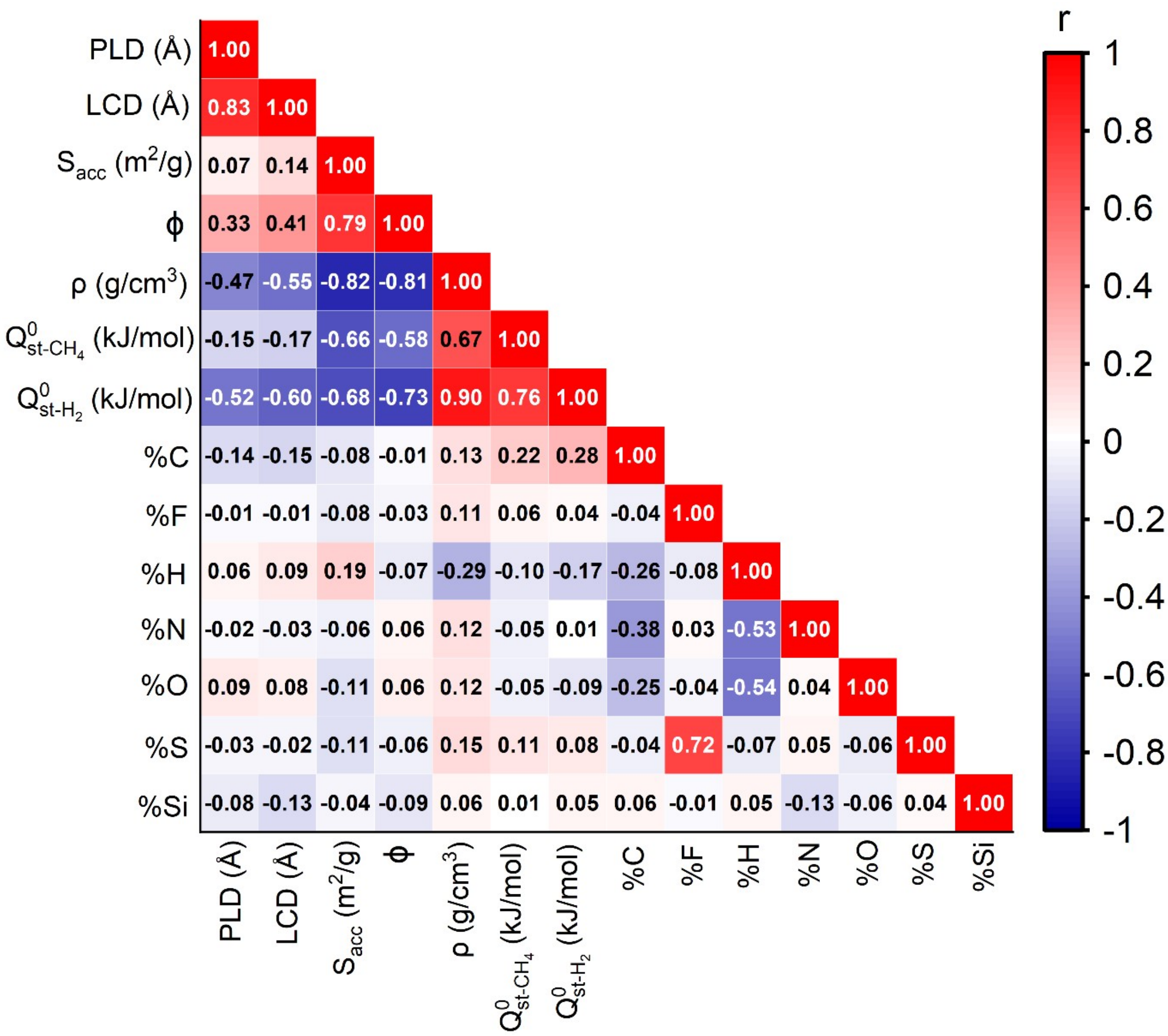

Figure S1. Correlation matrix for structural and chemical descriptors of 7,737 hypoCOFs. Pearson coefficients (r) are provided for the relationship between each descriptor.

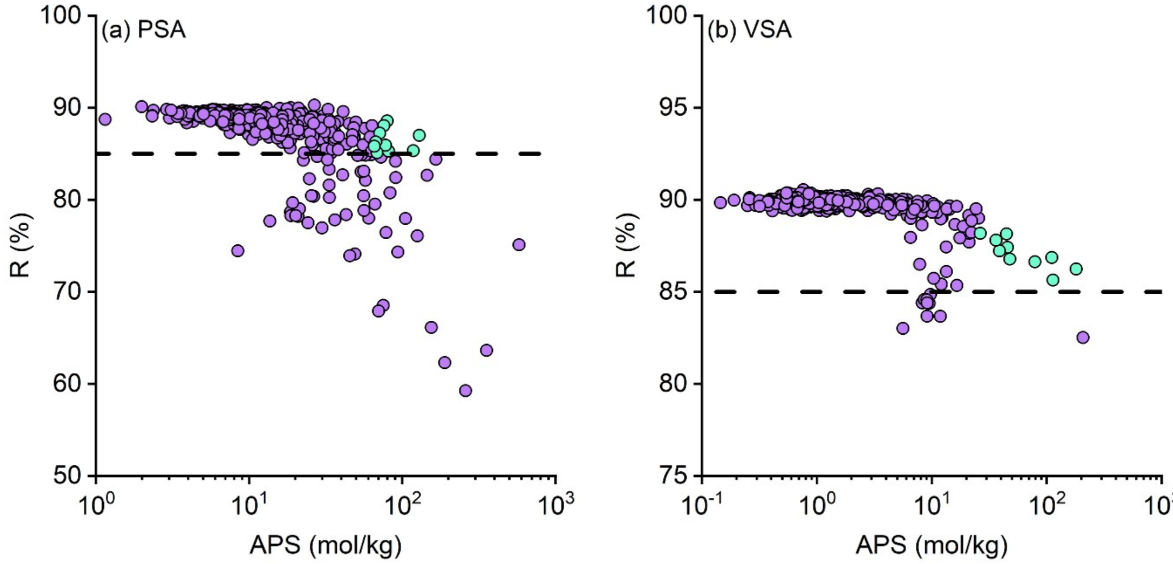

Figure S2. R% and APS of 597 CoRE COFs computed for CH_4/H_2:50/50 separation at (a) PSA and (b) VSA conditions.

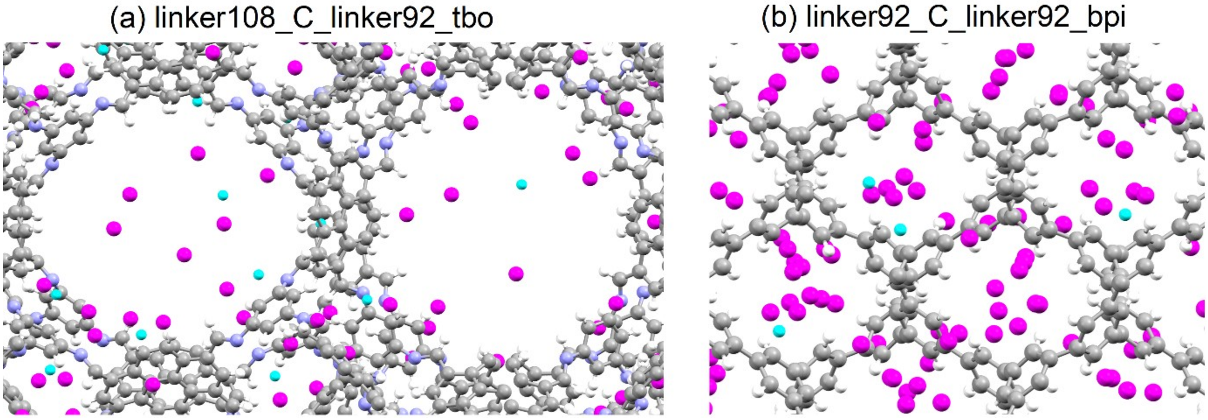

Figure S3. The snapshot showing the adsorption of the gas mixture in the best hypoCOFs identified at (a) PSA, (b) VSA processes at 10 bar, 298 K. CH_4 and H_2 molecules were denoted as magenta and cyan spheres, respectively. Grey and white, and light blue colored spheres represent C, H and N atoms in the framework.

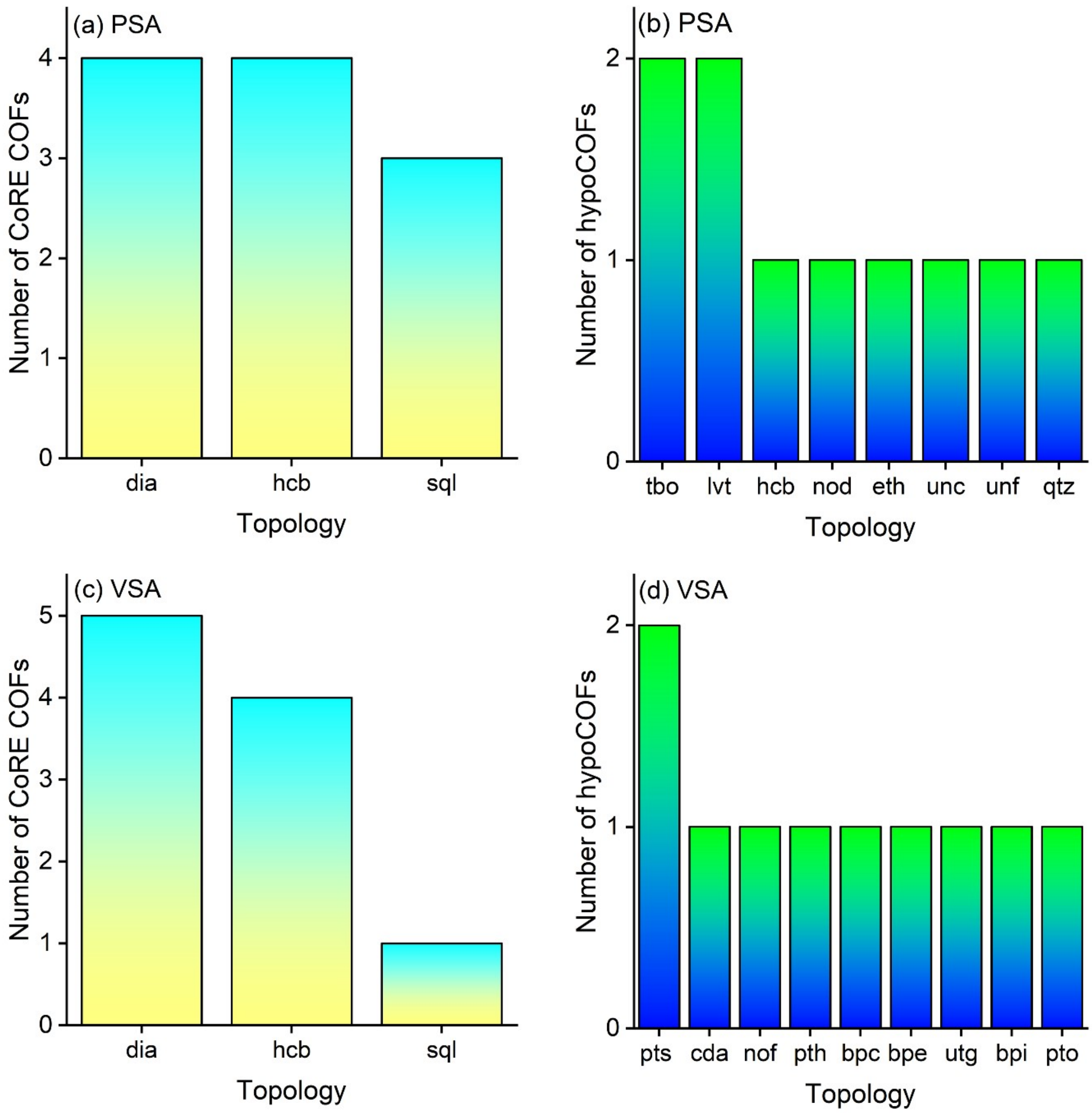

Figure S4. The topology distributions among the top 10 CoRE COFs and the top 10 hypoCOFs for CH_4/H_2 separation at (a-b) PSA and (c-d) VSA conditions.

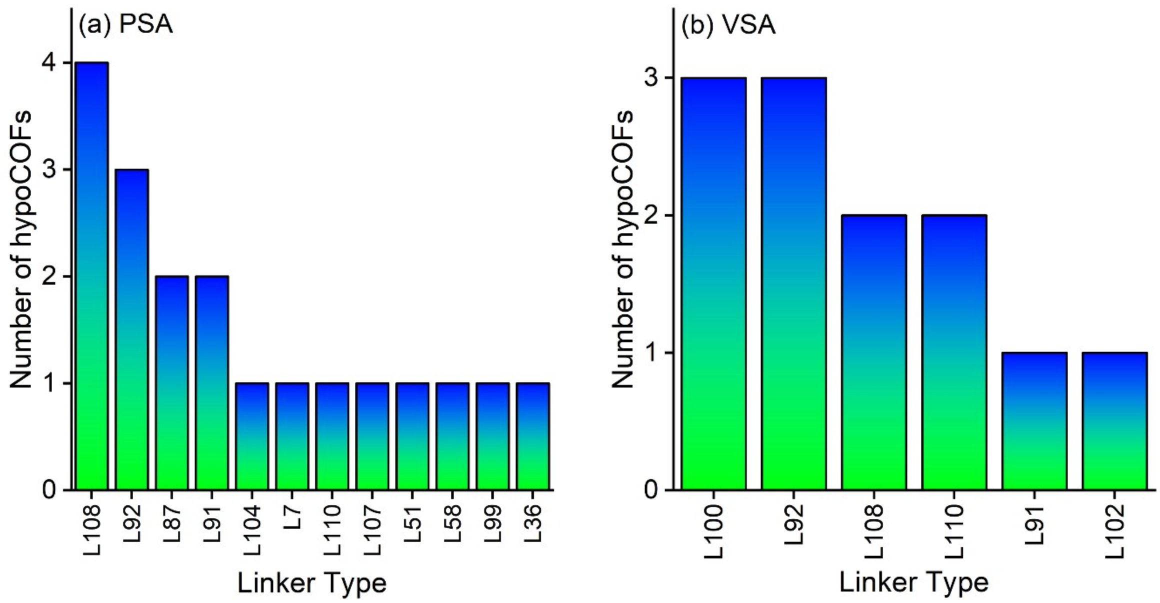

Figure S5. The distribution of linker types among the top 10 hypoCOFs for CH_4/H_2 separation at (a) PSA and (b) VSA conditions.

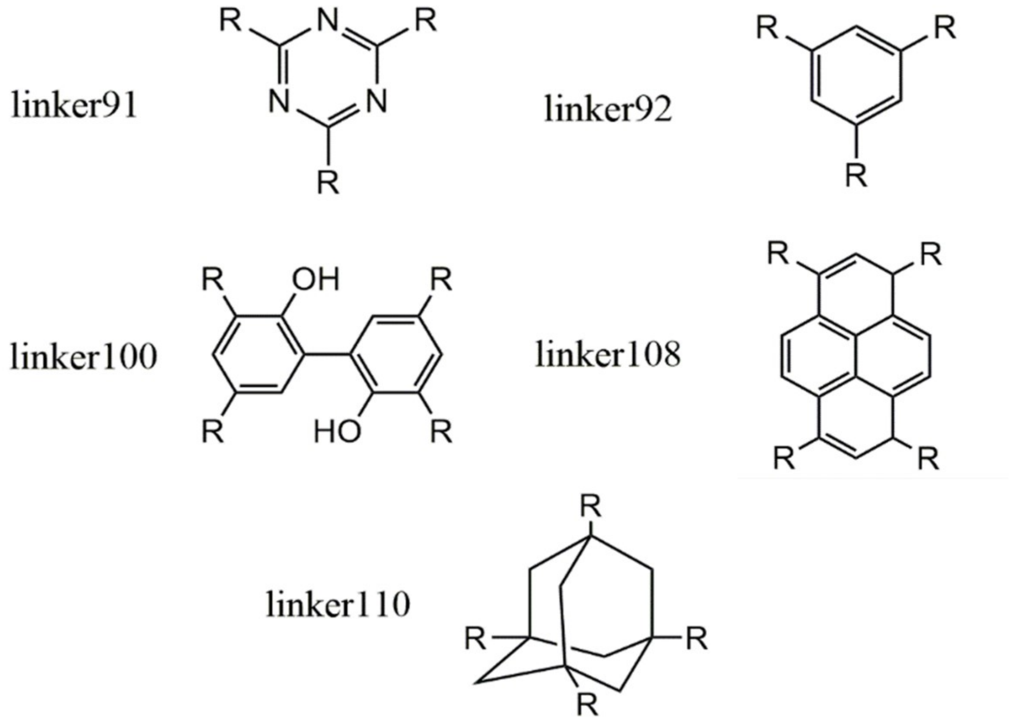

Figure S6. The schematic representations of the most frequent linkers found in top hypoCOFs. R represents the repeating unit of monomers.

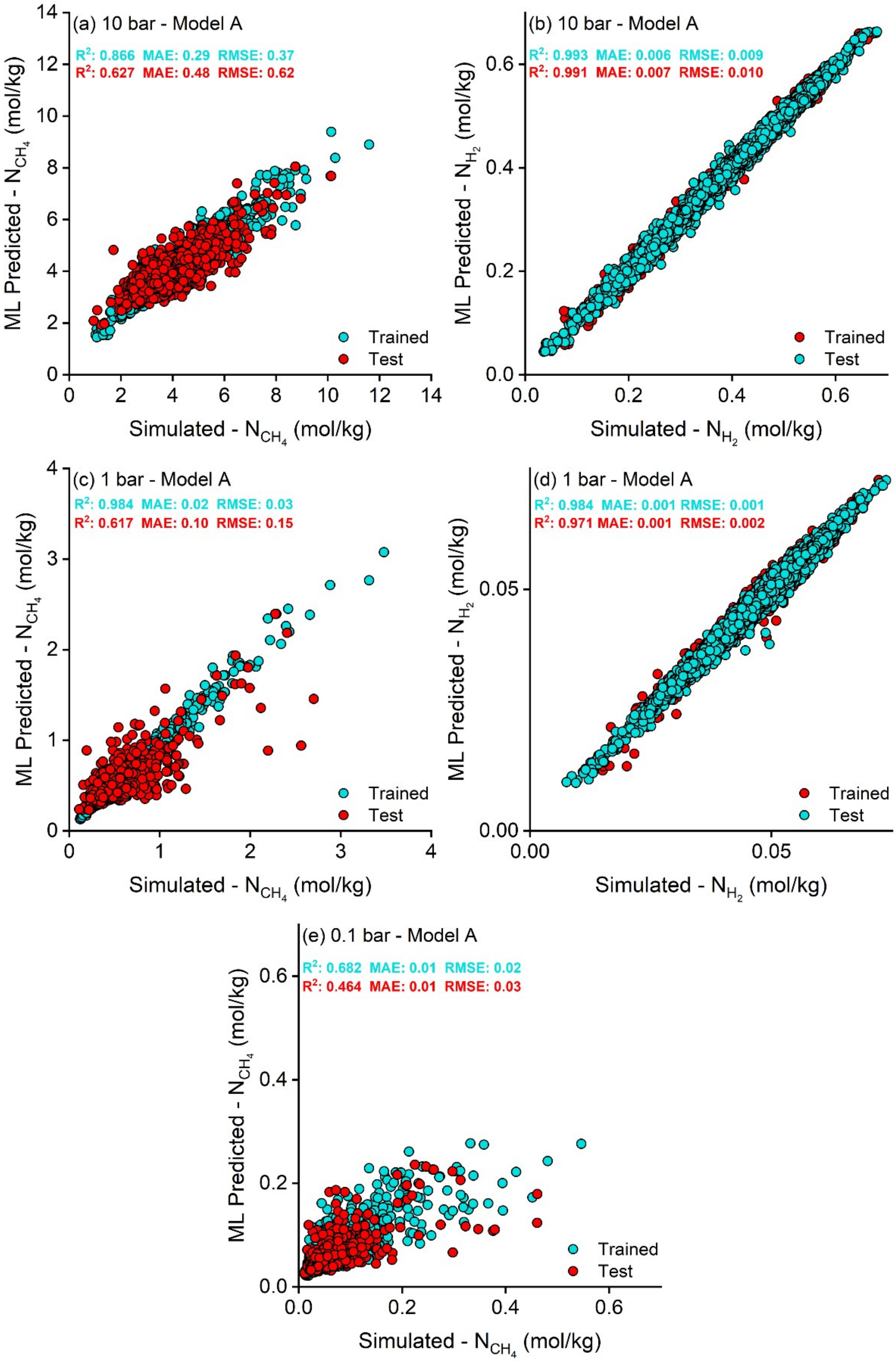

Figure S7. Comparison of CH_4 and H_2 uptakes predicted by ML models constructed with Group A descriptors and simulated uptakes in 7,737 hypoCOFs at (a-b) 10 bar, (c-d) 1 bar. Data at 0.1 bar is produced only for CH_4 for calculating the CH_4 working capacity at VSA condition and given in (e). Blue (red) symbols represent the training (test) data.

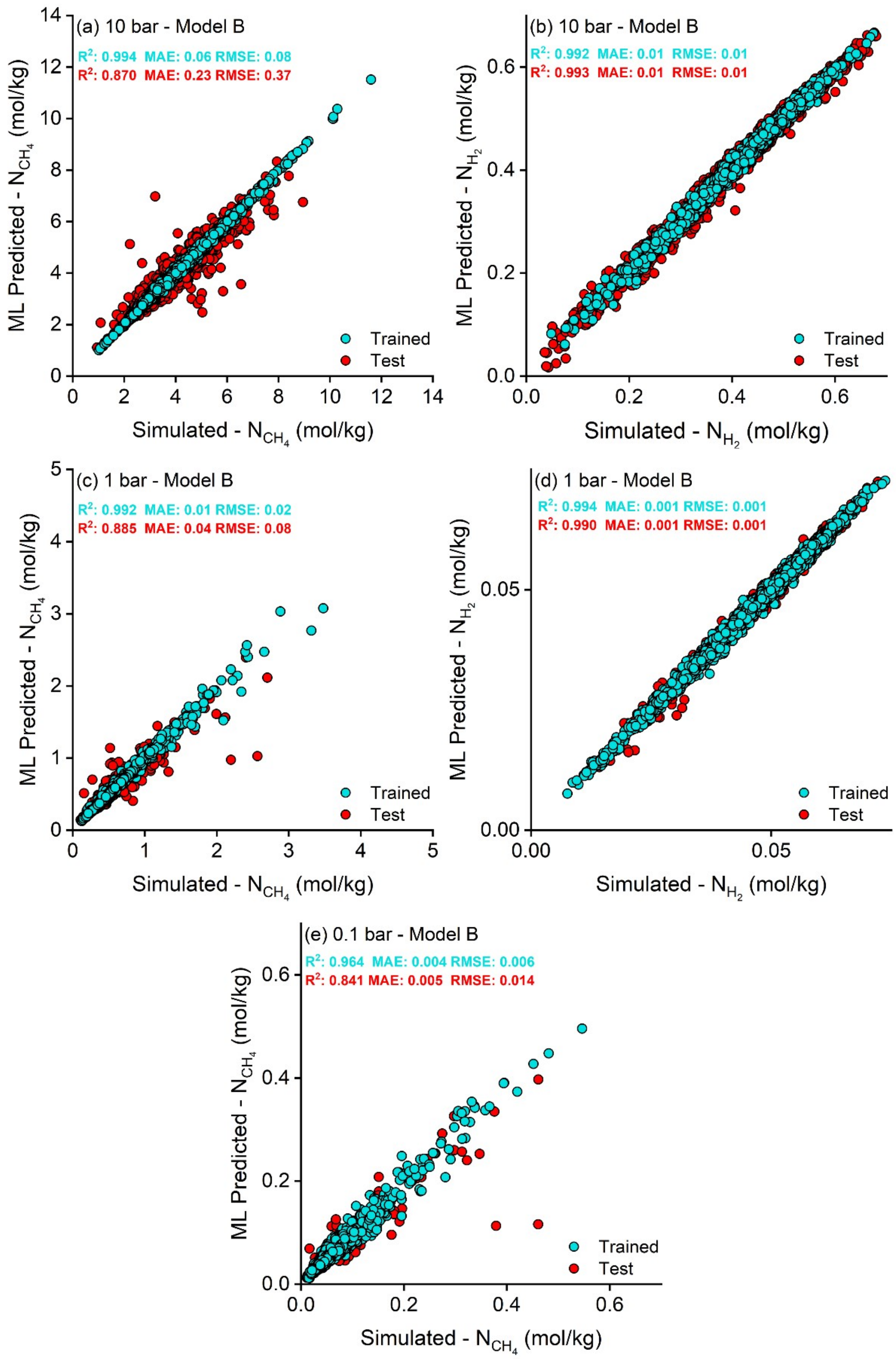

Figure S8. Comparison of CH_4 and H_2 uptakes predicted by ML models constructed with Group B descriptors and simulated uptakes in 7,737 hypoCOFs at (a-b) 10 bar, (c-d) 1 bar. Data at 0.1 bar is produced only for CH_4 for calculating the CH_4 working capacity at VSA condition and given in (e). Blue (red) symbols represent the training (test) data.

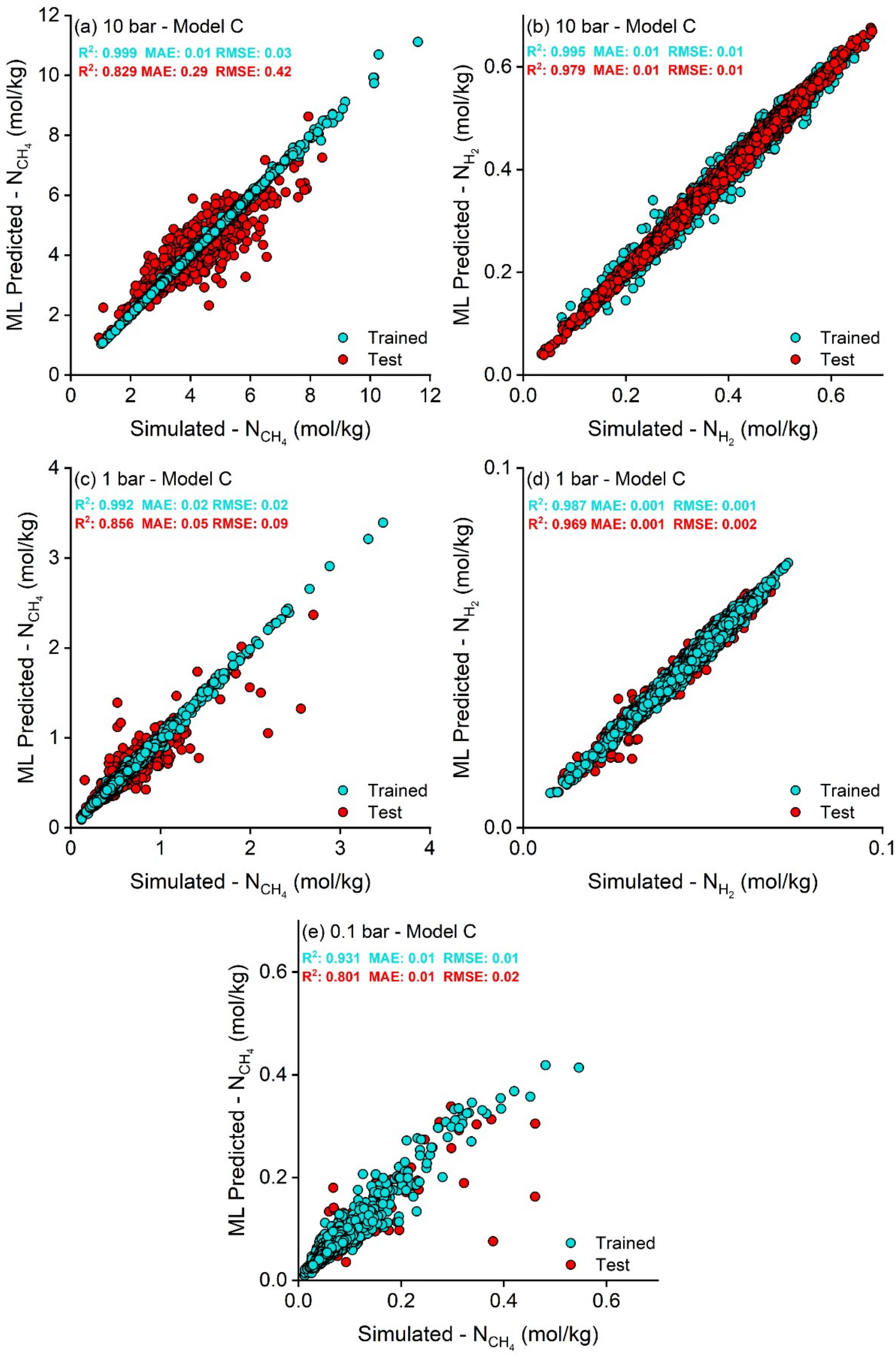

Figure S9. Comparison of CH_4 and H_2 uptakes predicted by ML models constructed with Group C descriptors and simulated uptakes in 7,737 hypoCOFs at (a-b) 10 bar, (c-d) 1 bar. Data at 0.1 bar is produced only for CH_4 for calculating the CH_4 working capacity at VSA condition and given in (e). Blue (red) symbols represent the training (test) data.

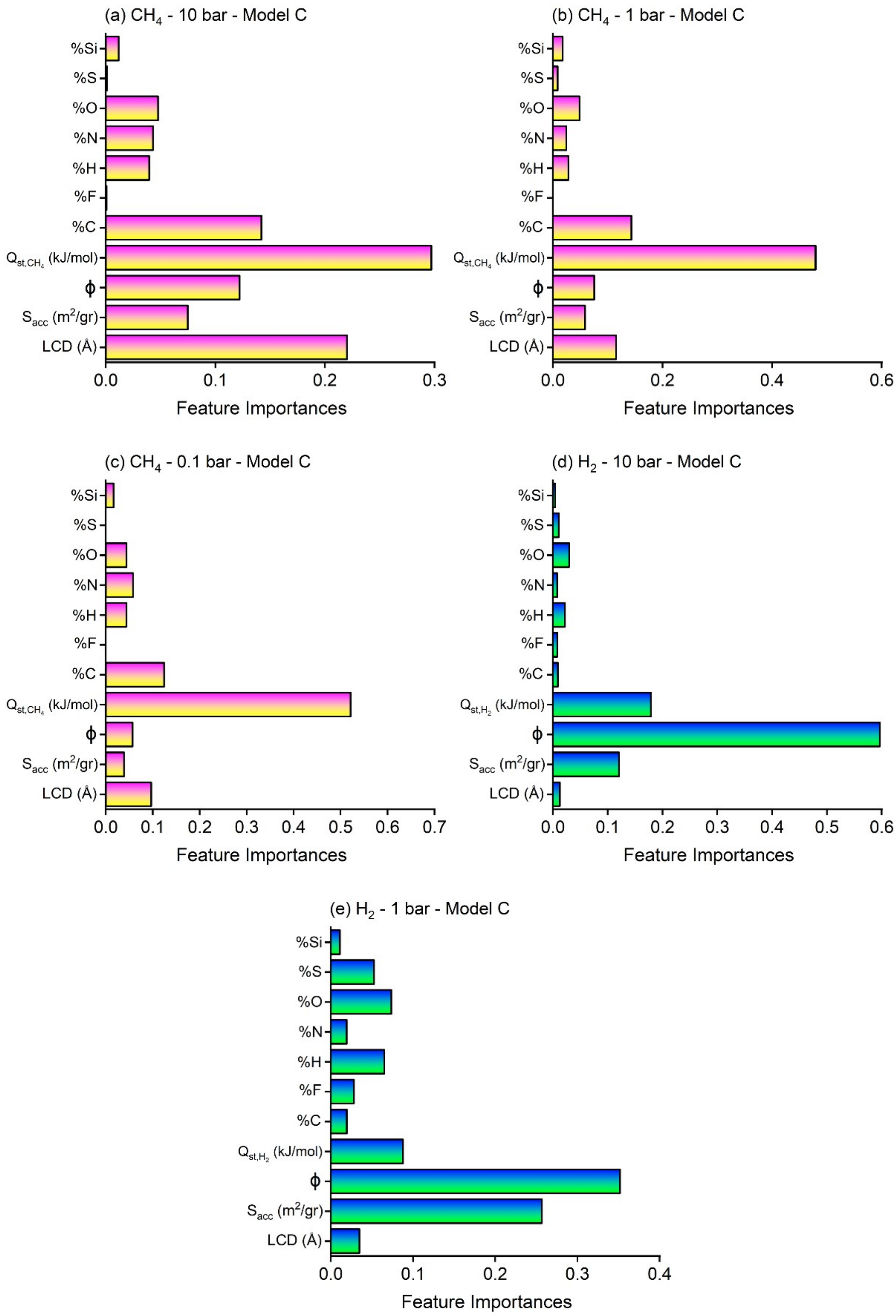

Figure S10. Feature importance distributions for Group C models used in predicting (a-c) CH_4 and (d,e) H_2 adsorption properties in 7,737 hypoCOFs at 0.1, 1, and 10 bar, 298 K.

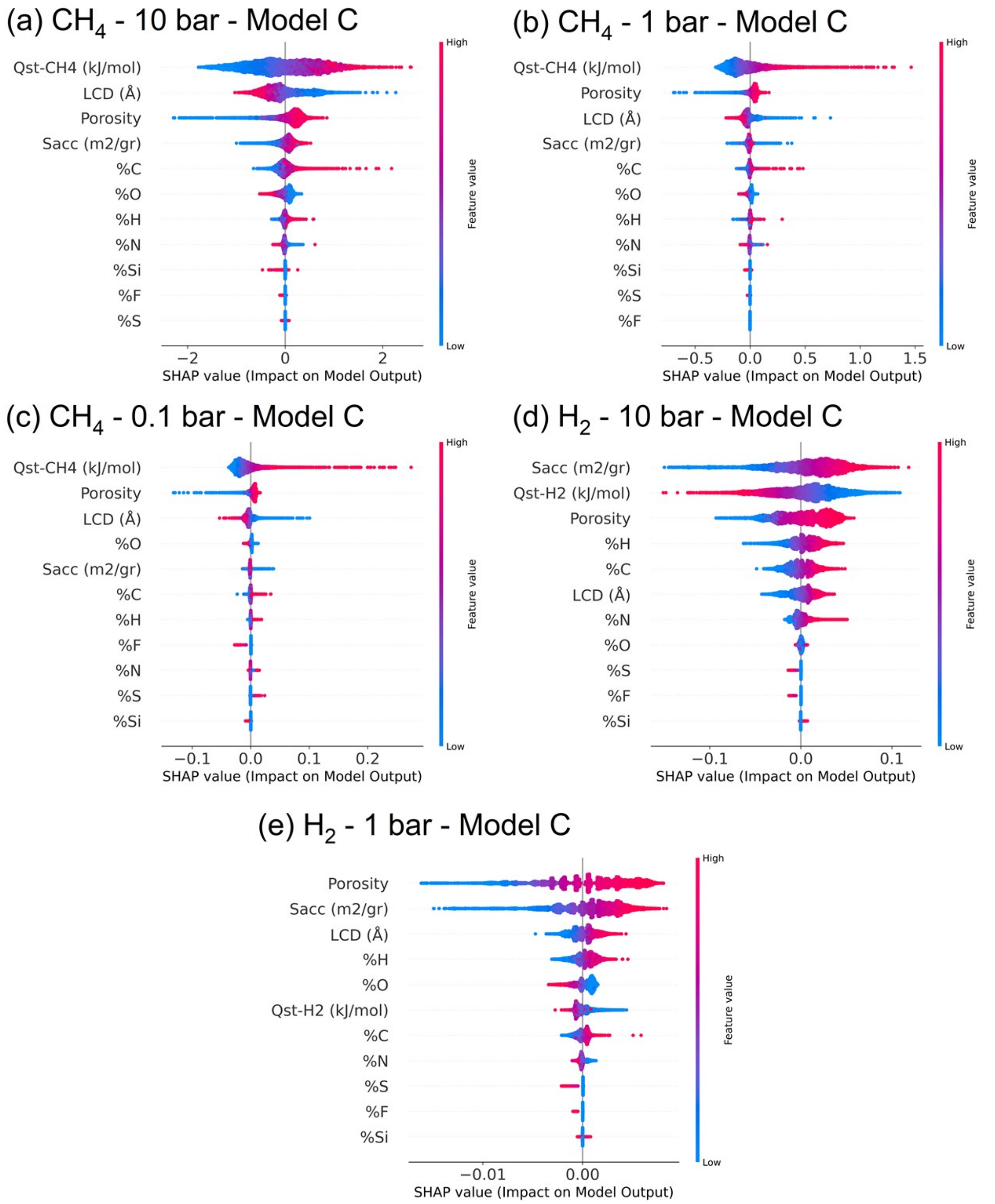

Figure S11. SHAP feature importance distributions for Group C models used in predicting CH_4 adsorption properties at (a) 10 bar, (b) 1 bar, (c) 0.1 bar, 298 K and H_2 adsorption properties at (d)10 bar, (e) 1 bar, 298 K. Features are sorted by their corresponding importances calculated by SHAP from top to bottom in descending order.

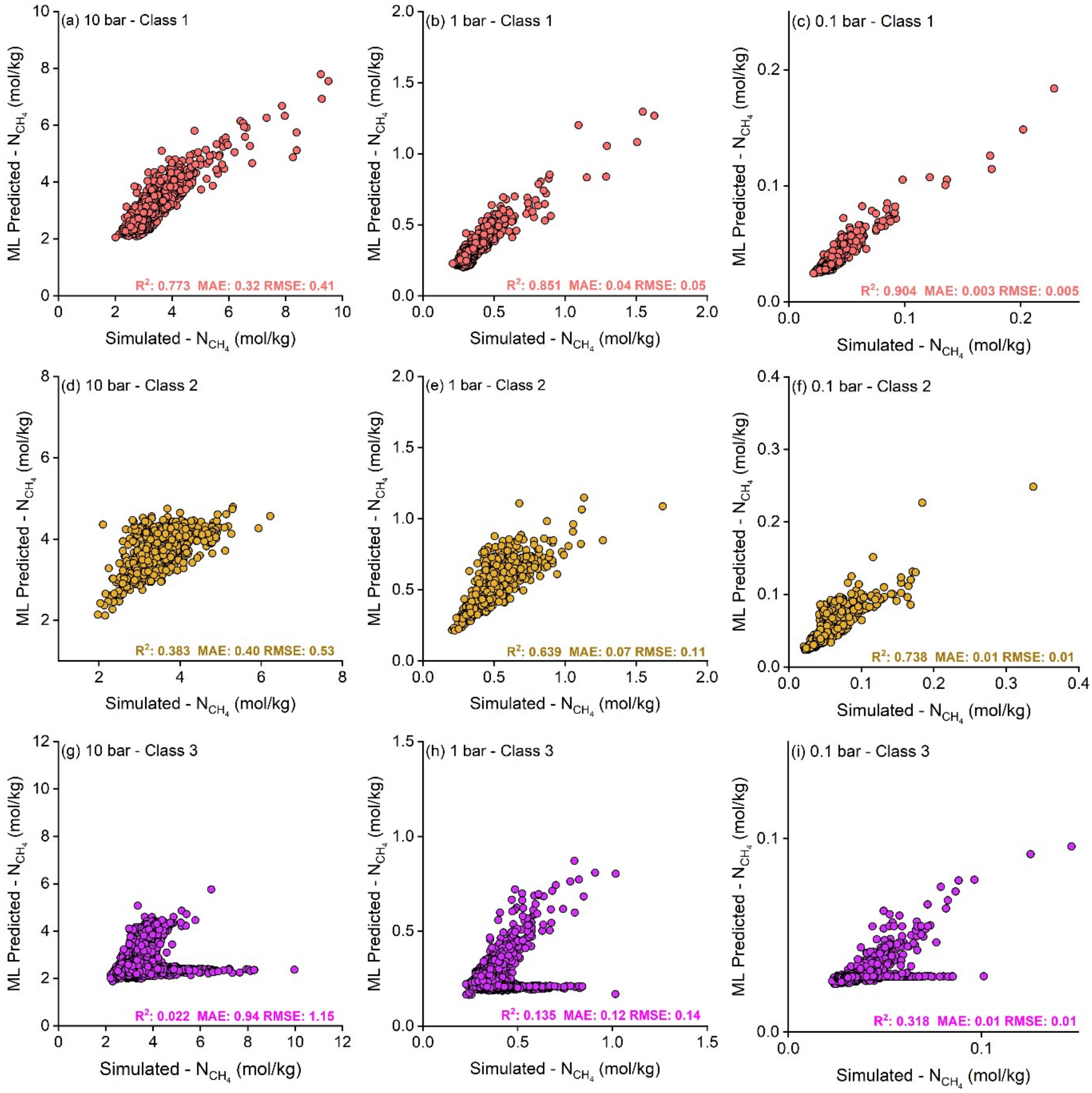

Figure S12. Comparison of CH_4 uptake data of unseen hypoCOFs predicted by Group C models at (a, d, g) 10, (b, e, h) 1, and (c, f, i) 0.1 bar.

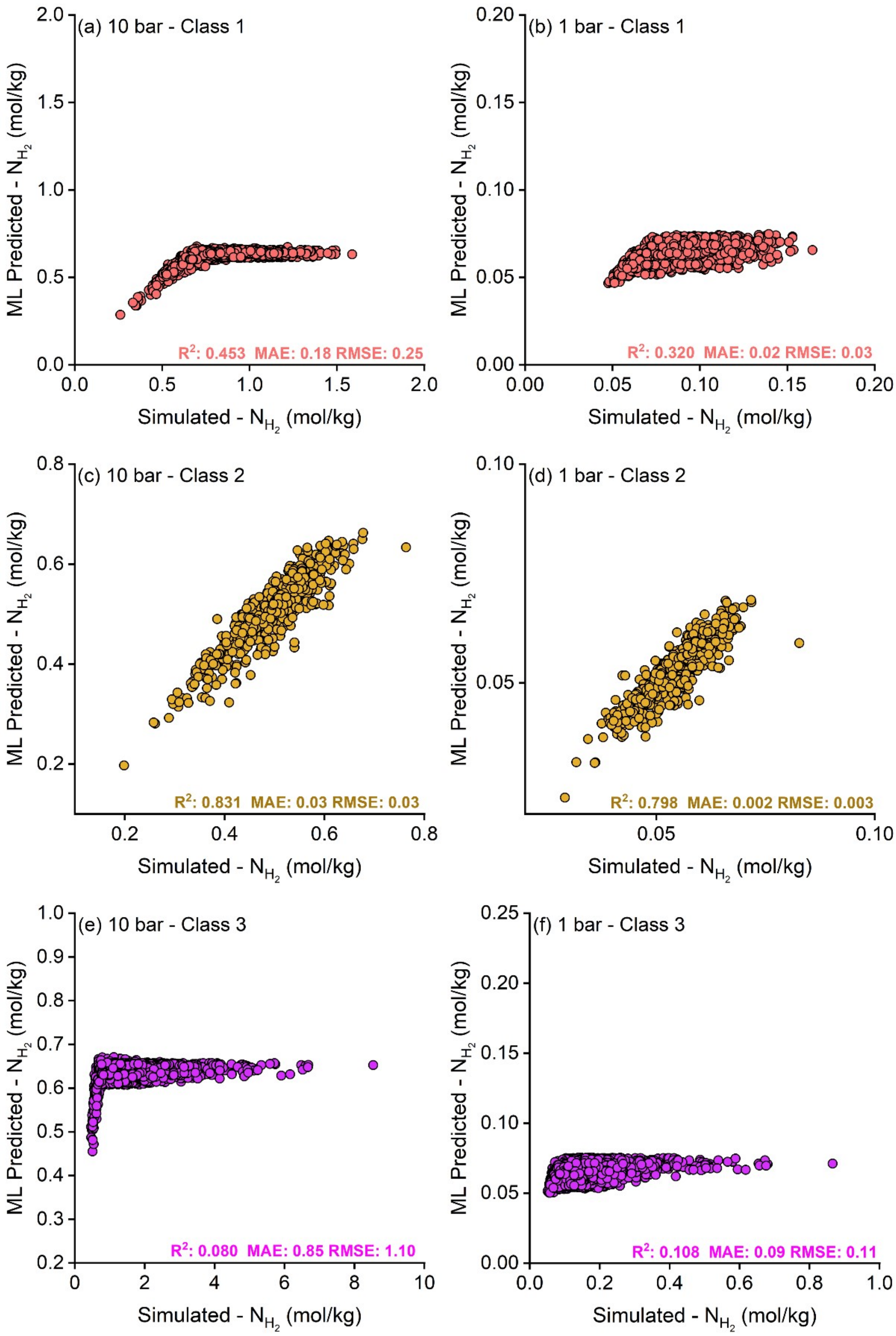

Figure S13. Comparison of H_2 uptake data of unseen hypoCOFs predicted by Group C models at (a, c, e) 10 bar and (b, d, f) 1 bar.

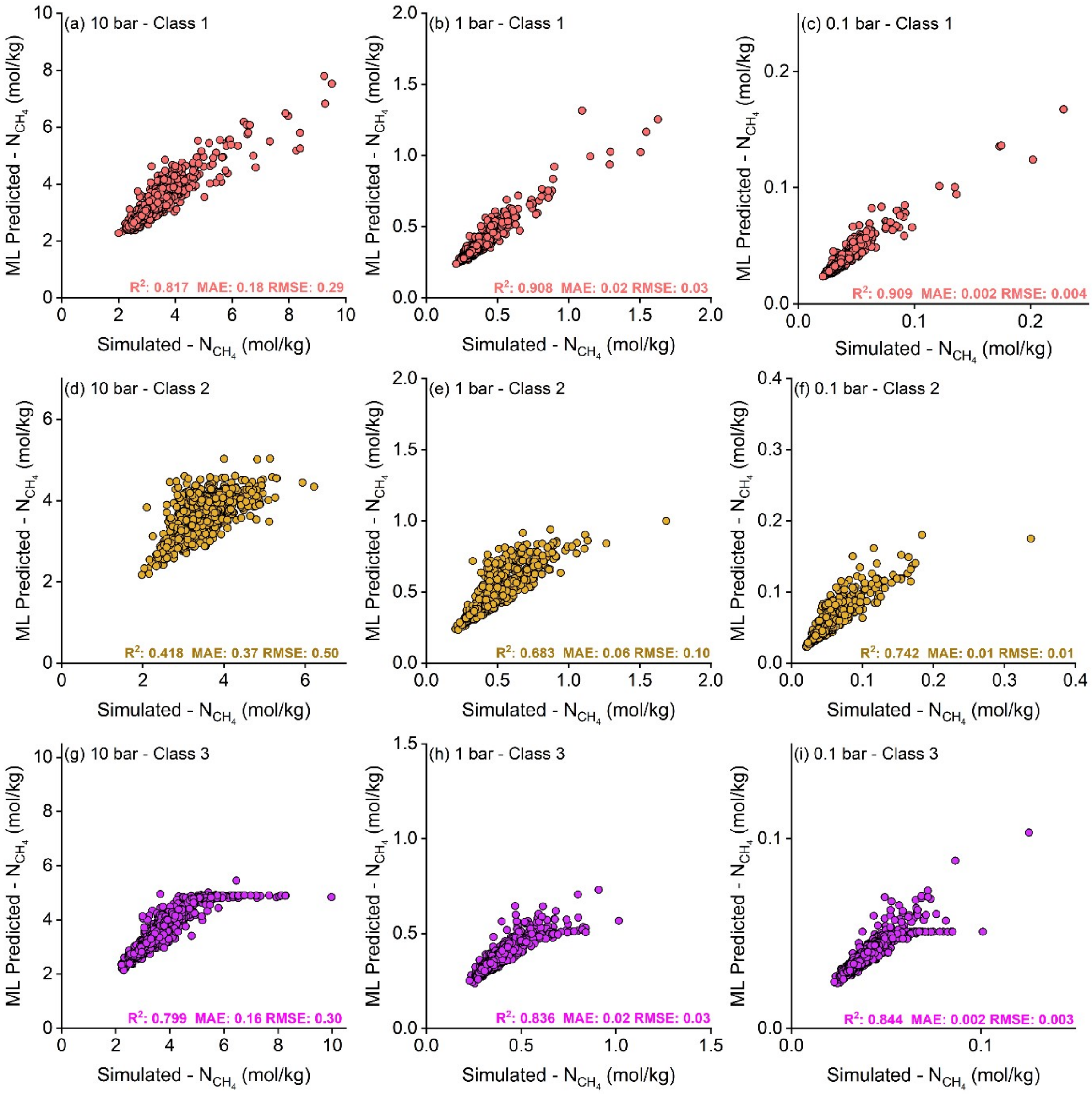

Figure S14. Comparison of CH_4 uptake data of unseen hypoCOFs predicted by extended Group C models at (a, d, g) 10, (b, e, h) 1, and (c, f, i) 0.1 bar.

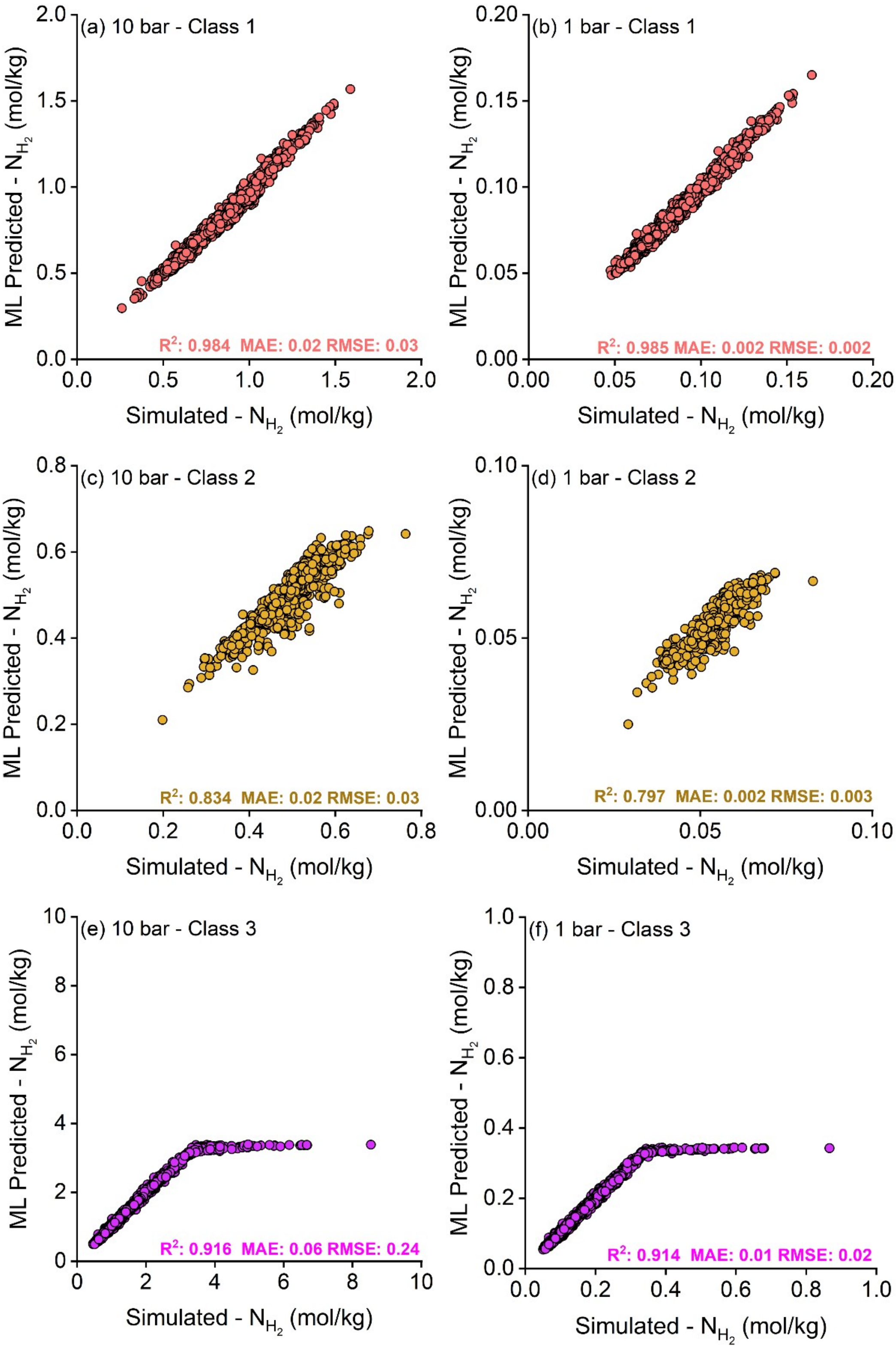

Figure S15. Comparison of H_2 uptake data of unseen hypoCOFs predicted by extended Group C models at (a, c, e) 10 bar, and (b, d, f) 1 bar.

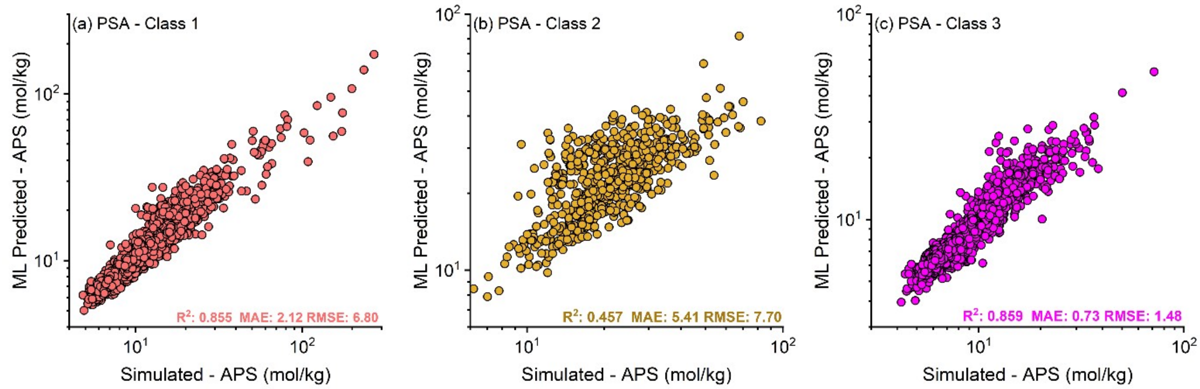

Figure S16. ML-predicted and simulated APSs for unseen hypoCOFs at PSA condition.

Table S1. Performance metrics computed to evaluate CoRE COF and hypoCOF adsorbents.

Metric	Formula
Mixture adsorption selectivity	$S_{CH_4/H_2} = \frac{N_{CH4}}{N_{H_2}} / \frac{y_{CH4}}{y_{H_2}}$
Working capacity (mol/kg)	$\Delta N_{CH4} = N_{ads,CH4} - N_{des,CH4}$
Adsorbent performance score (mol/kg)	$APS = S_{CH_4/H_2} \times \Delta N_{CH4}$
Percent regenerability	$R\% = \frac{\Delta N_{CH4}}{N_{ads,CH4}} \times 100\%$

$N_{ads,i}$: gas uptake at adsorption condition (mol/kg), $N_{des,i}$: gas uptake at desorption condition (mol/kg), y_i: bulk composition of the gas mixture

Table S2. The list of descriptors to construct machine learning models.

Group A	Group B	Group C
Largest Cavity Diameter (Å), LCD	Largest Cavity Diameter (Å), LCD	Largest Cavity Diameter (Å), LCD
Pore limiting diameter (Å), PLD	Pore limiting diameter (Å), PLD	Accessible Surface Area (m^2/g), S_{acc}
Accessible Surface Area (m^2/g), S_{acc}	Accessible Surface Area (m^2/g), S_{acc}	Porosity, φ
Porosity, φ	Porosity, φ	Heat of adsorption (kJ/mol), Q_{st}^0
Density (g/cm^3), ρ	Density (g/cm^3), ρ	Carbon percentage, C%
	Heat of adsorption (kJ/mol), Q_{st}^0	Hydrogen percentage, H%
	Carbon percentage, C%	Nitrogen percentage, N%
	Hydrogen percentage, H%	Oxygen percentage, O%
	Nitrogen percentage, N%	Fluorine percentage, F%
	Oxygen percentage, O%	Sulfur percentage, S%
	Fluorine percentage, F%	Silicon percentage, Si%
	Sulfur percentage, S%	
	Silicon percentage, Si%	

Table S3. The quantities calculated to evaluate the model accuracies.

Metric	Formula
Coefficient of Determination (R^2)	$1 - \frac{1}{M}\sum_{m=1}^{M}(\bar{y} - \hat{y})^2 / \frac{1}{M}\sum_{m=1}^{M}(y - \hat{y})^2$
Mean Absolute Error (MAE)	$\sum_{m=1}^{M}\lvert y - \hat{y}\rvert \Big/ M$
Root Mean Square Error (RMSE)	$\sqrt{\sum_{m=1}^{M}(y - \hat{y})^2 \Big/ M}$

M: the number of samples, y: simulated value, $\hat{y}$: predicted value, $\bar{y}$: average of the simulated value

Table S4. The ML pipelines constructed by Group A descriptors and their parameters based on the gas adsorption properties of COFs at each adsorption condition.

Property	Best Pipeline with Parameters
10 bar, 298 K (CH_4)	ExtraTreesRegressor(FastICA(PolynomialFeatures(input_matrix, degree=2, include_bias=False, interaction_only=False), tol=0.6000000000000001), bootstrap=False, max_features=0.45, min_samples_leaf=1, min_samples_split=13, n_estimators=100)
1 bar, 298 K (CH_4)	ExtraTreesRegressor(PolynomialFeatures(MaxAbsScaler(input_matrix), degree=2, include_bias=False, interaction_only=False), bootstrap=False, max_features=0.45, min_samples_leaf=1, min_samples_split=4, n_estimators=100)
0.1 bar, 298 K (CH_4)	RandomForestRegressor(ElasticNetCV(GradientBoostingRegressor(PCA(input_matrix , iterated_power=7, svd_solver=randomized), alpha=0.8, learning_rate=0.001, loss=quantile, max_depth=5, max_features=0.5, min_samples_leaf=15, min_samples_split=11, n_estimators=100, subsample=0.2), l1_ratio=0.25, tol=0.01), bootstrap=True, max_features=0.7000000000000001, min_samples_leaf=7, min_samples_split=8, n_estimators=100)
10 bar, 298 K (H_2)	XGBRegressor(LassoLarsCV(DecisionTreeRegressor(PolynomialFeatures(input_matri x, degree=2, include_bias=False, interaction_only=False), max_depth=2, min_samples_leaf=7, min_samples_split=18), normalize=False), learning_rate=0.1, max_depth=4, min_child_weight=10, n_estimators=100, n_jobs=1, objective=reg:squarederror, subsample=0.7500000000000001, verbosity=0)
1 bar, 298 K (H_2)	XGBRegressor(CombineDFs(SelectFwe(input_matrix, alpha=0.045), PCA(input_matrix, iterated_power=10, svd_solver=randomized)), learning_rate=0.1, max_depth=9, min_child_weight=10, n_estimators=100, n_jobs=1, objective=reg:squarederror, subsample=0.7500000000000001, verbosity=0)

Table S5. The ML pipelines constructed by Group B descriptors and their parameters based on the gas adsorption properties of COFs at each adsorption condition.

Property	Best Pipeline with Parameters
10 bar, 298 K (CH_4)	GradientBoostingRegressor(LassoLarsCV(input_matrix, normalize=False), alpha=0.8, learning_rate=0.1, loss=ls, max_depth=9, max_features=0.9000000000000001, min_samples_leaf=1, min_samples_split=5, n_estimators=100, subsample=0.8500000000000001)
1 bar, 298 K (CH_4)	ExtraTreesRegressor(CombineDFs(RidgeCV(input_matrix), input_matrix), bootstrap=False, max_features=0.9000000000000001, min_samples_leaf=2, min_samples_split=2, n_estimators=100)
0.1 bar, 298 K (CH_4)	XGBRegressor(RidgeCV(PCA(input_matrix, iterated_power=7, svd_solver=randomized)), learning_rate=0.1, max_depth=4, min_child_weight=7, n_estimators=100, n_jobs=1, objective=reg:squarederror, subsample=0.9000000000000001, verbosity=0)
10 bar, 298 K (H_2)	LassoLarsCV(PolynomialFeatures(RobustScaler(input_matrix), degree=2, include_bias=False, interaction_only=False), normalize=False)
1 bar, 298 K (H_2)	GradientBoostingRegressor(LassoLarsCV(PolynomialFeatures(input_matrix, degree=2, include_bias=False, interaction_only=False), normalize=False), alpha=0.8, learning_rate=0.1, loss=ls, max_depth=5, max_features=0.9000000000000001, min_samples_leaf=1, min_samples_split=9, n_estimators=100, subsample=0.25)

Table S6. The ML pipelines constructed by Group C descriptors and their parameters based on the gas adsorption properties of COFs at each adsorption condition.

Property	Best Pipeline with Parameters
10 bar, 298 K (CH_4)	ExtraTreesRegressor(PolynomialFeatures(input_matrix, degree=2, include_bias=False, interaction_only=False), bootstrap=False, max_features=0.55, min_samples_leaf=1, min_samples_split=3, n_estimators=100)
1 bar, 298 K (CH_4)	XGBRegressor(ExtraTreesRegressor(PolynomialFeatures(SGDRegressor(input_matrix , alpha=0.001, eta0=1.0, fit_intercept=False, l1_ratio=0.0, learning_rate=constant, loss=huber, penalty=elasticnet, power_t=0.5), degree=2, include_bias=False, interaction_only=False), bootstrap=False, max_features=0.8, min_samples_leaf=5, min_samples_split=17, n_estimators=100), learning_rate=0.5, max_depth=4, min_child_weight=14, n_estimators=100, n_jobs=1, objective=reg:squarederror, subsample=1.0, verbosity=0)
0.1 bar, 298 K (CH_4)	XGBRegressor(LassoLarsCV(input_matrix, normalize=False), learning_rate=0.1, max_depth=4, min_child_weight=14, n_estimators=100, n_jobs=1, objective=reg:squarederror, subsample=1.0, verbosity=0)
10 bar, 298 K (H_2)	XGBRegressor(LassoLarsCV(SelectFwe(input_matrix, alpha=0.04), normalize=False), learning_rate=0.1, max_depth=8, min_child_weight=2, n_estimators=100, n_jobs=1, objective=reg:squarederror, subsample=0.6500000000000001, verbosity=0)
1 bar, 298 K (H_2)	XGBRegressor(SGDRegressor(input_matrix, alpha=0.01, eta0=0.01, fit_intercept=True, l1_ratio=1.0, learning_rate=constant, loss=huber, penalty=elasticnet, power_t=0.0), learning_rate=0.1, max_depth=8, min_child_weight=2, n_estimators=100, n_jobs=1, objective=reg:squarederror, subsample=0.6500000000000001, verbosity=0)

Table S7. Separation performances of the top 10 CoRE COFs for PSA process.

ID	Dimension	APS (mol/kg)	S_{CH_4/H_2}	R%	LCD (Å)	Density (g/cm^3)	ϕ	S_{acc} (m^2/g)
114	3D	129.1	28.1	87.0	8.98	0.74	0.63	2412.2
468	2D	117.9	27.2	85.2	5.58	0.74	0.63	2412.2
133	2D	81.2	30.0	88.8	9.15	1.00	0.69	2899.2
11	3D	79.6	14.2	89.4	6.33	0.46	0.72	3521.4
5	2D	77.8	27.1	88.8	8.25	0.83	0.68	3449.7
329	2D	75.7	16.2	88.9	6.69	0.59	0.68	3262.3
294	3D	71.3	17.0	85.8	8.66	0.62	0.56	2137.7
69	2D	68.3	25.9	87.3	8.48	0.84	0.56	1793.5
435	3D	67.0	13.0	87.4	16.68	0.47	0.60	2700.7
140	2D	65.6	13.4	89.5	12.53	0.46	0.68	3306.4

Table S8. Separation performances of the top 10 CoRE COFs for VSA process.

ID	Dimension	APS (mol/kg)	S_{CH_4/H_2}	R%	LCD (Å)	Density (g/cm^3)	ϕ	S_{acc} (m^2/g)
327	3D	180.6	105.4	86.2	5.27	0.98	0.54	1125.0
113	3D	113.2	136.0	85.6	4.27	1.12	0.43	280.3
328	3D	110.9	93.6	86.9	6.15	1.05	0.50	940.7
116	3D	78.9	108.7	86.6	4.51	1.04	0.46	425.1
414	2D	47.7	43.2	86.8	14.93	0.87	0.69	1495.8
193	2D	45.6	60.5	87.4	6.74	1.03	0.52	1045.3
76	3D	44.6	52.7	88.1	5.66	0.89	0.55	621.3
480	2D	38.8	72.1	87.2	4.73	1.25	0.44	548.6
380	2D	36.2	38.2	87.8	7.34	0.70	0.64	2401.2
431	2D	26.4	45.1	88.2	6.89	0.91	0.53	1142.0

Table S9. Separation performances of the top 10 hypoCOFs for PSA process.

ID	APS (mol/kg)	S_{CH_4/H_2}	R%	LCD (Å)	Density (g/cm^3)	ϕ	S_{acc} (m^2/g)
linker108_C_linker92_C_tbo_relaxed	205.4	28.7	86.9	15.14	0.61	0.74	2813.0
linker108_C_linker91_C_tbo_relaxed	181.4	26.6	86.5	15.44	0.64	0.78	2533.2
linker92_C_linker92_C_nod_relaxed	177.1	24.6	86.1	7.70	0.56	0.78	4098.8
linker92_C_linker92_C_eth_relaxed	166.4	21.9	88.8	9.71	0.53	0.80	3869.6
linker107_C_linker7_C_unc_relaxed	162.6	27.7	85.7	6.09	0.60	0.61	3952.4
linker110_C_linker87_C_unf_relaxed	159.6	24.1	87.4	10.19	0.55	0.70	4663.4
linker91_C_linker51_C_hcb_relaxed	150.6	25.4	86.4	7.14	0.62	0.70	4294.7
linker108_C_linker87_C_lvt_relaxed _interp_2	150.2	23.2	85.0	10.07	0.58	0.72	2944.6
linker108_C_linker58_C_lvt_relaxed _interp_2	147.6	22.6	87.0	10.13	0.60	0.76	3209.6
linker99_C_linker36_C_qtz_relaxed_ interp_2	146.8	21.9	87.4	8.36	0.53	0.77	4497.8

Table S10. Separation performances of the top 10 hypoCOFs for VSA process.

ID	APS (mol/kg)	S_{CH_4/H_2}	R%	LCD (Å)	Density (g/cm^3)	ϕ	S_{acc} (m^2/g)
linker92_C_linker92_C_bpi_relaxed	242.9	79.4	87.9	5.52	0.71	0.74	3012.7
linker110_C_linker100_C_pth_relaxed	234.0	112.2	86.3	4.92	0.92	0.50	1266.9
linker91_C_linker91_C_nof_relaxed_interp_2	232.1	128.6	85.2	5.65	1.13	0.39	719.7
linker108_C_linker100_C_pts_relaxed	182.1	85.6	87.5	5.88	0.83	0.60	1803.8
linker100_C_linker108_C_pts_relaxed	176.5	83.7	87.7	5.88	0.83	0.60	1803.8
linker92_C_linker92_C_utg_relaxed	170.0	66.0	89.2	5.16	0.69	0.71	3416.4
linker92_C_linker92_C_bpe_relaxed	169.1	82.2	85.9	5.45	0.79	0.60	1382.6
linker92_C_linker92_C_bpc_relaxed	151.8	64.7	88.2	6.49	0.70	0.71	2619.6
linker100_C_linker102_C_cda_relaxed	149.6	92.3	85.6	6.60	0.87	0.47	1301.0
linker110_C_linker87_C_mdf_relaxed	149.0	119.0	85.5	6.04	0.96	0.25	588.8

Table S11. The correlations between ML-predicted and simulated values of training and test sets for given models and conditions. R^2 is a unitless number between 0 and 1, as both MAE and RMSE have units of mol/kg.

Model	Train Set	Test Set
A - CH_4 at 10 bar	R^2:0.866, MAE:0.285, RMSE:0.371	R^2:0.627, MAE:0.477, RMSE:0.621
A - CH_4 at 1 bar	R^2:0.984, MAE:0.016, RMSE:0.029	R^2:0.617, MAE:0.096, RMSE:0.149
A - CH_4 at 0.1 bar	R^2:0.682, MAE:0.010, RMSE:0.018	R^2:0.464, MAE:0.013, RMSE:0.026
A – H_2 at 10 bar	R^2:0.993, MAE:0.006, RMSE:0.009	R^2:0.991, MAE:0.007, RMSE:0.010
A – H_2 at 1 bar	R^2:0.984, MAE:0.001, RMSE:0.001	R^2:0.971, MAE:0.001, RMSE:0.002
B - CH_4 at 10 bar	R^2:0.994, MAE:0.061, RMSE:0.079	R^2:0.870 MAE:0.232, RMSE:0.367
B - CH_4 at 1 bar	R^2:0.992, MAE:0.008, RMSE:0.020	R^2:0.885, MAE:0.037, RMSE:0.082
B - CH_4 at 0.1 bar	R^2:0.964, MAE:0.004, RMSE:0.006	R^2:0.841, MAE:0.005, RMSE:0.014
B – H_2 at 10 bar	R^2:0.992, MAE:0.006, RMSE:0.009	R^2:0.993, MAE:0.006, RMSE:0.009
B – H_2 at 1 bar	R^2:0.994, MAE:0.001, RMSE:0.001	R^2:0.990, MAE:0.001, RMSE:0.001
C - CH_4 at 10 bar	R^2:0.999, MAE:0.014, RMSE:0.027	R^2:0.829, MAE:0.290, RMSE:0.421
C - CH_4 at 1 bar	R^2:0.992, MAE:0.015, RMSE:0.020	R^2:0.856, MAE:0.047, RMSE:0.091
C - CH_4 at 0.1 bar	R^2:0.931, MAE:0.005, RMSE:0.008	R^2:0.801, MAE:0.006, RMSE:0.016
C – H_2 at 10 bar	R^2:0.995, MAE:0.005, RMSE:0.007	R^2:0.979, MAE:0.011, RMSE:0.015
C – H_2 at 1 bar	R^2:0.987, MAE:0.001, RMSE:0.001	R^2:0.969, MAE:0.001 RMSE:0.001

.

Table S12. The extended ML pipelines constructed by Group C descriptors and their parameters based on the gas adsorption properties of COFs at each adsorption condition.

Property	Best Pipeline with Parameters
10 bar, 298 K (CH_4)	RandomForestRegressor(RandomForestRegressor(ExtraTreesRegressor(ZeroCount(PolynomialFeatures(input_matrix, degree=2, include_bias=False, interaction_only=False)), bootstrap=False, max_features=0.8, min_samples_leaf=5, min_samples_split=17, n_estimators=100), bootstrap=False, max_features=0.2, min_samples_leaf=20, min_samples_split=20, n_estimators=100), bootstrap=False, max_features=0.2, min_samples_leaf=6, min_samples_split=4, n_estimators=100)
1 bar, 298 K (CH_4)	ExtraTreesRegressor(LassoLarsCV(PolynomialFeatures(input_matrix, degree=2, include_bias=False, interaction_only=False), normalize=False), bootstrap=False, max_features=0.55, min_samples_leaf=1, min_samples_split=16, n_estimators=100)
0.1 bar, 298 K (CH_4)	ExtraTreesRegressor(MaxAbsScaler(PolynomialFeatures(input_matrix, degree=2, include_bias=False, interaction_only=False)), bootstrap=False, max_features=0.9500000000000001, min_samples_leaf=3, min_samples_split=4, n_estimators=100)
10 bar, 298 K (H_2)	ExtraTreesRegressor(input_matrix, bootstrap=False, max_features=0.9500000000000001, min_samples_leaf=1, min_samples_split=4, n_estimators=100)
1 bar, 298 K (H_2)	ExtraTreesRegressor(input_matrix, bootstrap=False, max_features=0.9500000000000001, min_samples_leaf=1, min_samples_split=4, n_estimators=100)