XGBoost: An Optimal Machine Learning Model with just Structural Features to Discover MOF Adsorbents of Xe/Kr

Supporting Information

Heng Liang[1], Kun Jiang[2], Tong-an Yan[3] and Guang-hui Chen[1]*

1 Department of Chemistry, Key Laboratory for Preparation and Application of Ordered Structural Materials of Guangdong Province, Shantou University, Shantou 515063, Guangdong Province, China

2 Department of Natural Science, Shantou Polytechnic, Shantou 515041, Guangdong Province, China

3 State Key Laboratory of Organic-Inorganic Composites, Beijing University of Chemical Technology, Beijing 100029, China

E-mail: ghchen@stu.edu.cn

Table of Contents

Table S1. **The optimum regression model of Xe/Kr adsorption separation on MOFs in testing set at 1 bar and 298 K, where MSE, MAE, RMSE and R^2 represents mean square error, mean absolute error and determination coefficient, respectively. For adsorption capacity, the unit of MSE, MAE and RMSE is $mmol^2/g^2$, mmol/g and mmol/g, respectively.**

property	model	MSE	MAE	RMSE	R^2
$N_{(Xe)}$	Ridge	0.035	0.101	0.187	0.393
	LASSO	0.035	0.102	0.187	0.393
	Elastic net	0.035	0.102	0.187	0.392
	SVM	0.038	0.130	0.195	0.334
	Bayesian	0.035	0.101	0.187	0.392
	ANN	0.026	0.095	0.161	0.560
	RF	0.006	0.044	0.077	0.883
	XGBoost	0.003	0.029	0.055	0.951
$S_{(Xe/Kr)}$	Ridge	0.761	0.506	0.872	0.688
	LASSO	0.762	0.506	0.873	0.686
	Elastic net	0.762	0.507	0.873	0.687
	SVM	0.827	0.432	0.909	0.660
	Bayesian	0.760	0.505	0.872	0.687
	ANN	0.411	0.352	0.641	0.831
	RF	0.164	0.229	0.405	0.933
	XGBoost	0.065	0.147	0.255	0.973

Table S2. Cluster analysis of three categories of MOFs in Class A, B and C. Note that the N_{xe} (mmol/g) and $S_{Xe/Kr}$ are average values.

	MSE	NOS	N_{xe}	MSE	NOS	$S_{Xe/Kr}$
A	0.459	476	2.016	8.35	1542	9.696
B	0.104	376	2.492	30.76	10	17.892
C	0.298	586	1.2	1.763	15281	6.094

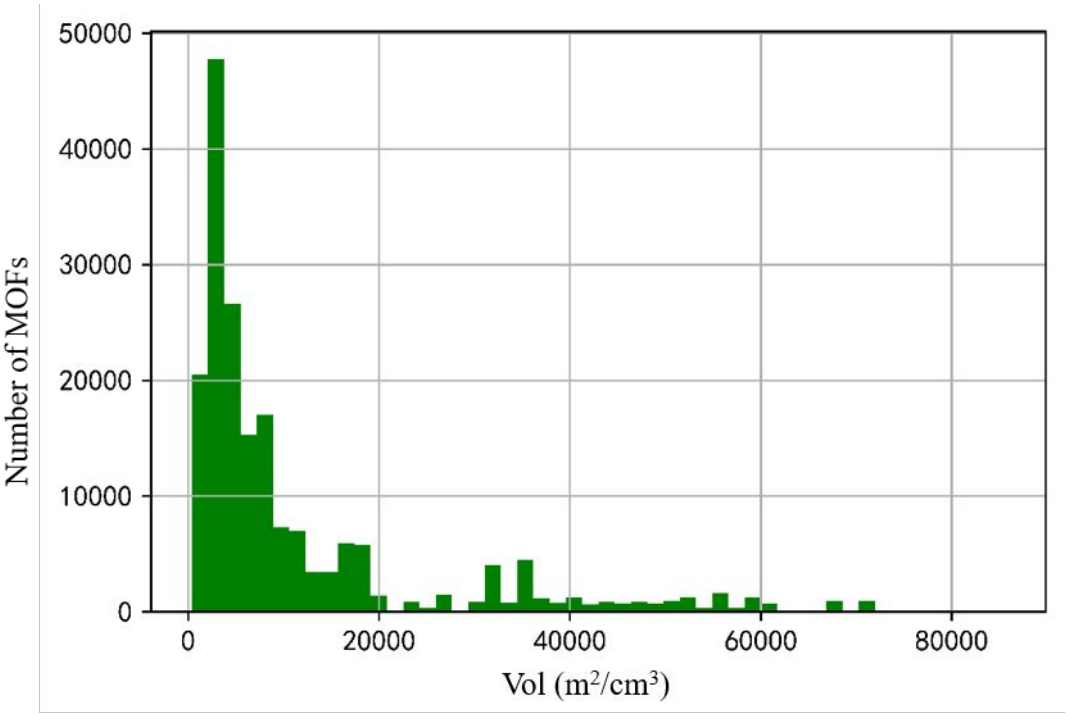
NOS: Number of Samples

Table S3. The statistics analysis of CO_2/CH_4 selectivity (sco_{2/CH_4}) and adsorption uptake (N_{CO_2}) (mmol/g) of MOFs in G-MOFs database at 298 K and 1 bar.

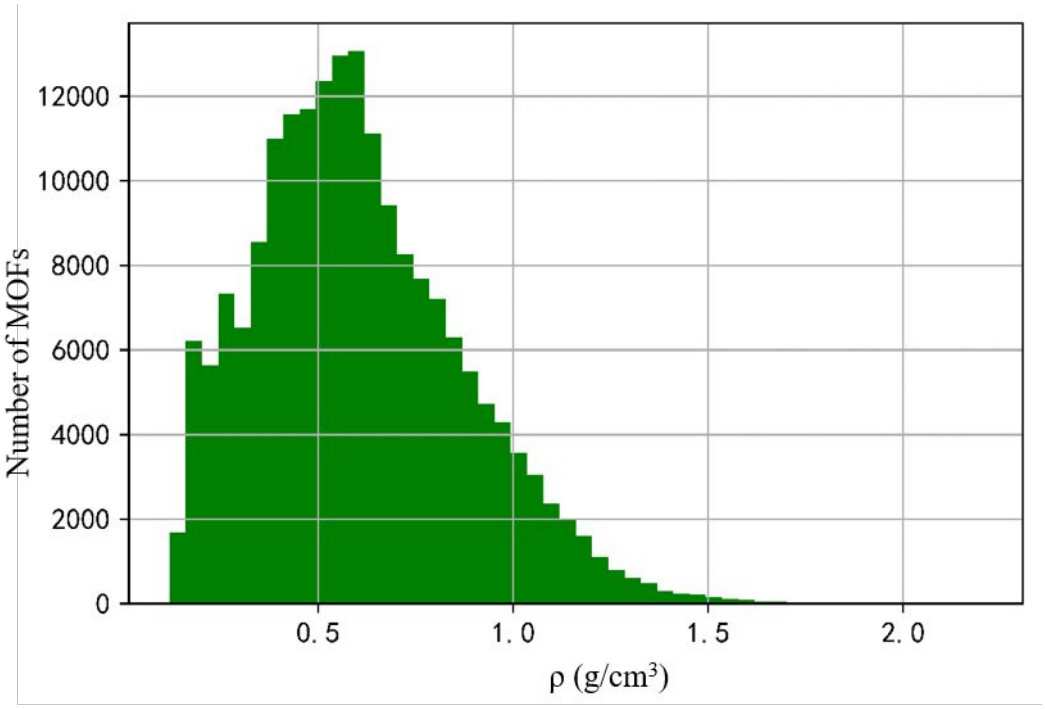
property / data analysis	sco_{2/CH_4}	N_{CO_2}
Std	2.5871	1.1336
50%	3.4023	1.0499
75%	4.7658	1.8968
Max	85.3023	11.9110

Table S4. The optimum regression models of CO_2/CH_4 adsorption separation on MOFs in testing set at 1 bar and 298 K, where R^2 represents determination coefficient.

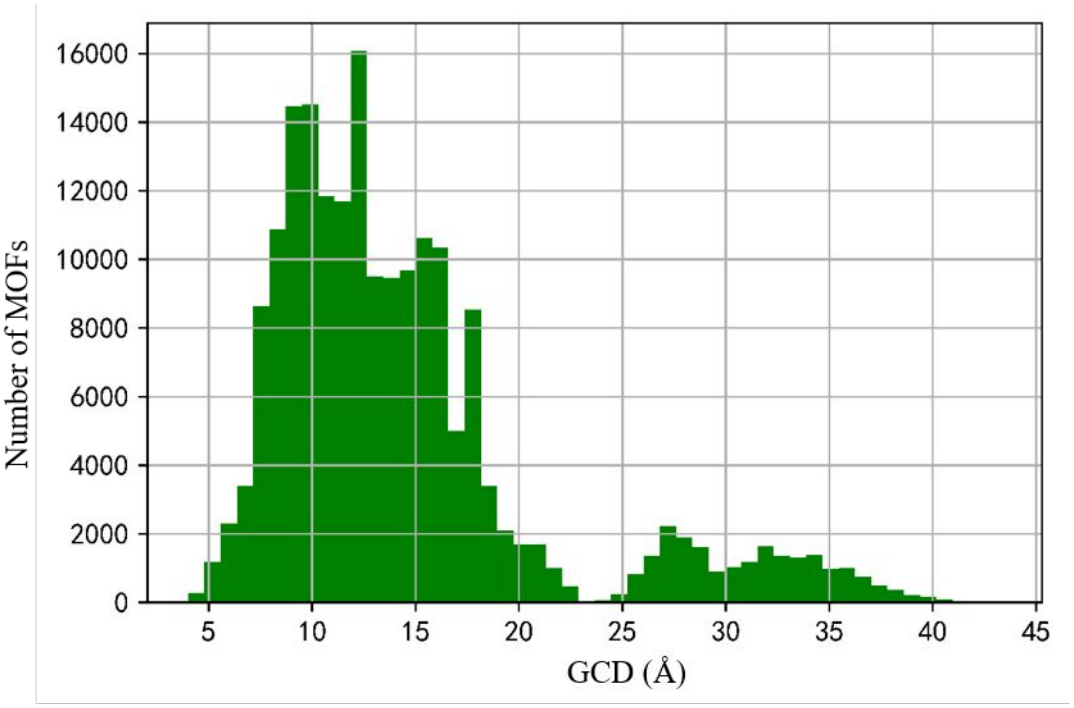
property / ML model	$R^2(sco_{2/CH_4})$	$R^2(N_{CO_2})$
Ridge	0.2246	0.4007
LASSO	0.2247	0.4000
Elastic net	0.2245	0.4001
SVM	0.2126	0.4864
Bayesian	0.2245	0.4006
ANN	0.3382	0.6413
RF	0.6461	0.8391
XGBoost	0.6836	0.8817


50000
40000
30000
20000
10000
0
Number of MOFs
0
20000
40000
60000
80000
Vol (m²/cm³)

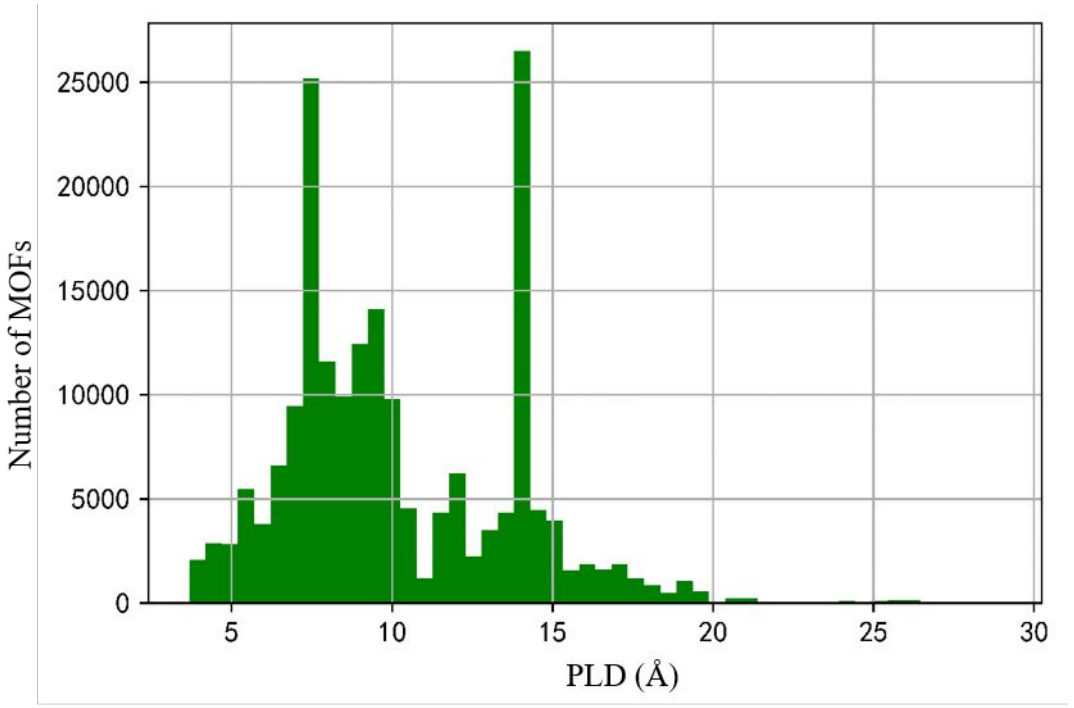
(a)


12000
10000
8000
6000
4000
2000
0
Number of MOFs
0.5
1.0
1.5
2.0
ρ (g/cm³)

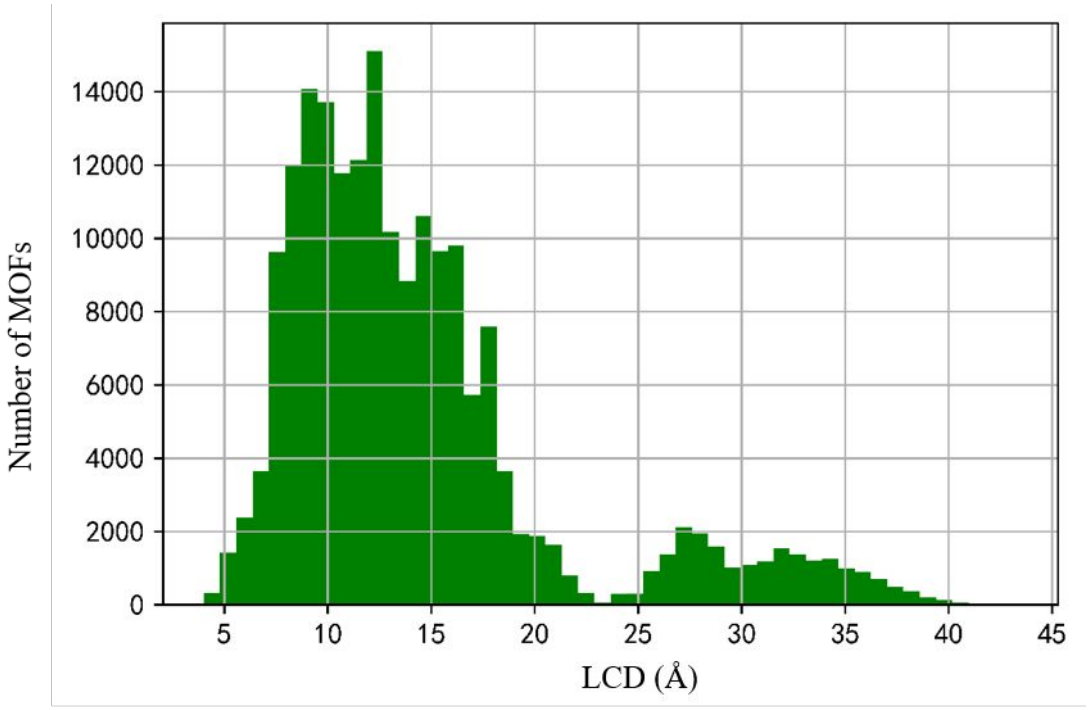
(b)


Number of MOFs
16000
14000
12000
10000
8000
6000
4000
2000
0
5
10
15
20
25
30
35
40
45
GCD (Å)

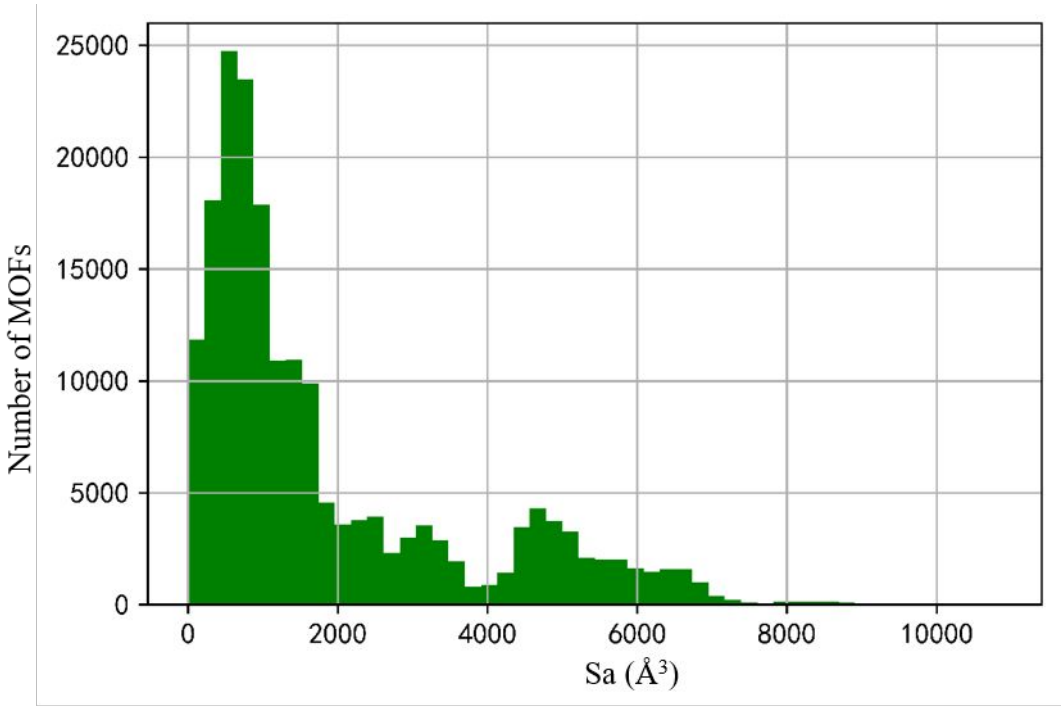
(c)


Number of MOFs
25000
20000
15000
10000
5000
0
5
10
15
20
25
30
PLD (Å)

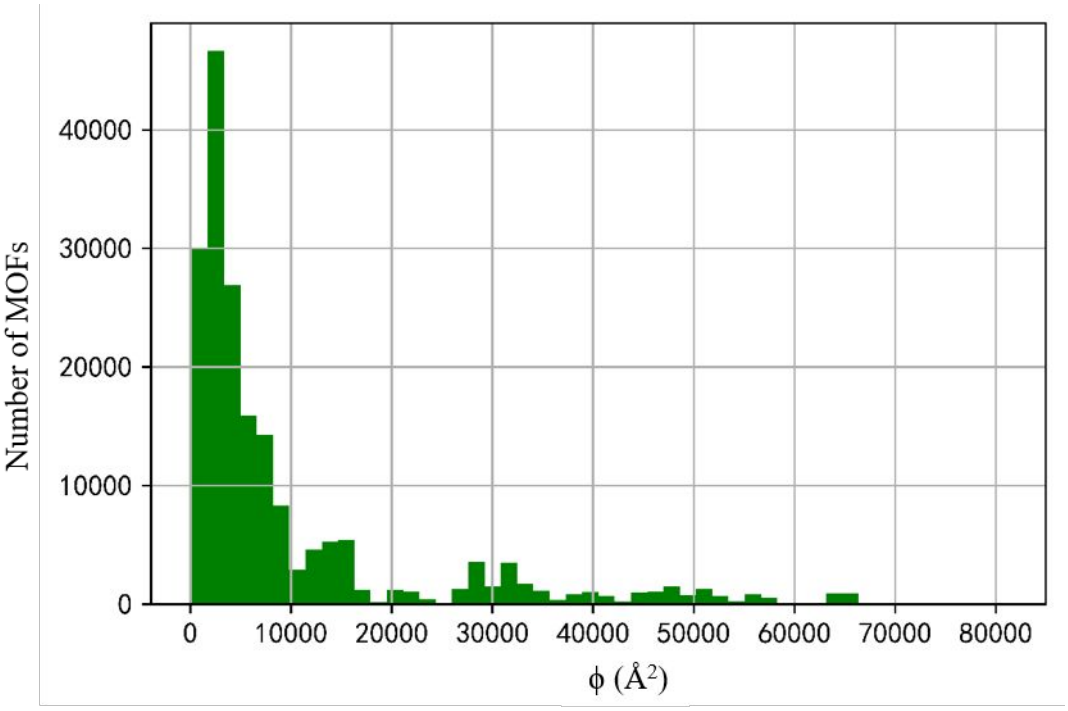
(d)


Number of MOFs
14000
12000
10000
8000
6000
4000
2000
0
5
10
15
20
25
30
35
40
45
LCD (Å)

(e)


Number of MOFs
25000
20000
15000
10000
5000
0
0
2000
4000
6000
8000
10000
Sa (Å³)

(f)



(g)

Figure S1. The plotted histogram of adsorption property distribution of G-MOFs upon: a) pore volume (Vol); b) density (ρ); c) global cavity diameter (GCD); d) pore limiting diameter (PLD); e) large cavity diameter (LCD); f) specific surface area (Sa) and g) porosity (ϕ).

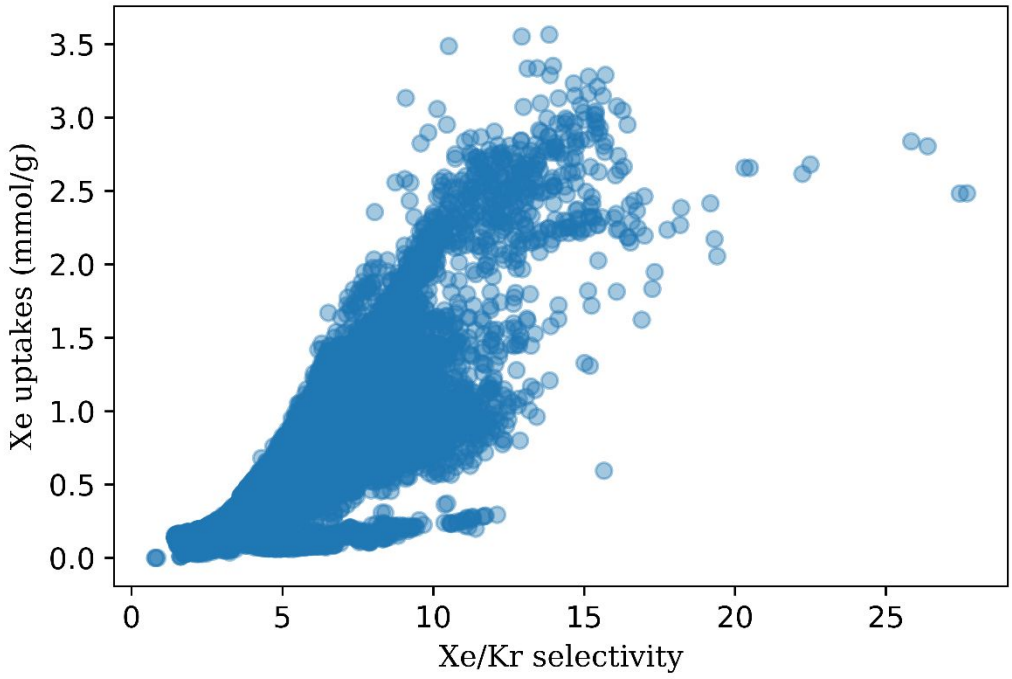
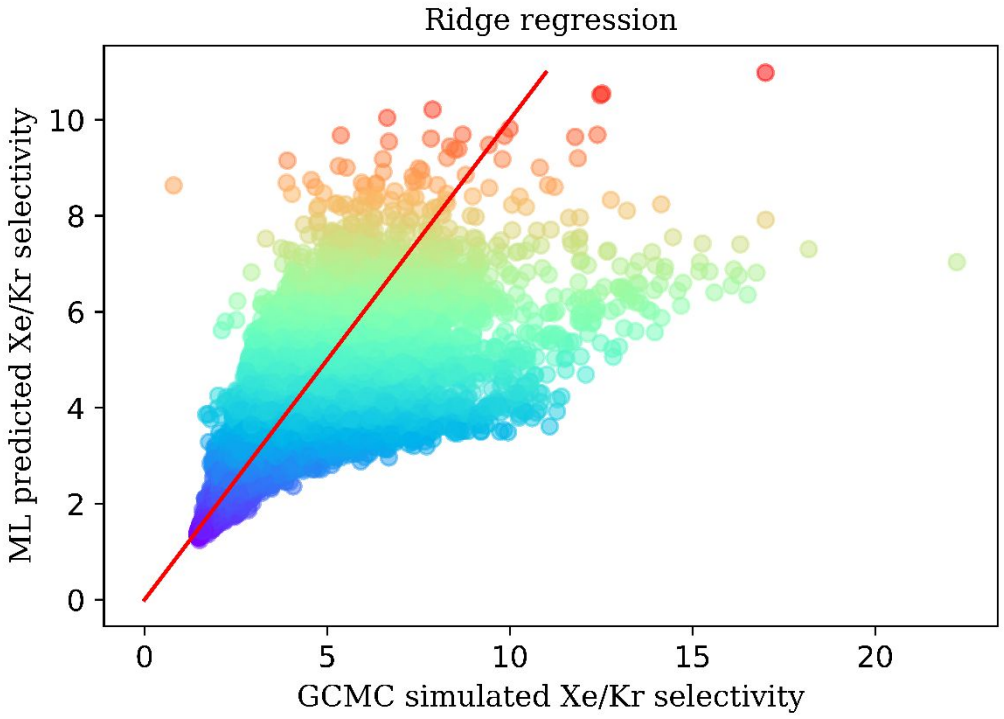
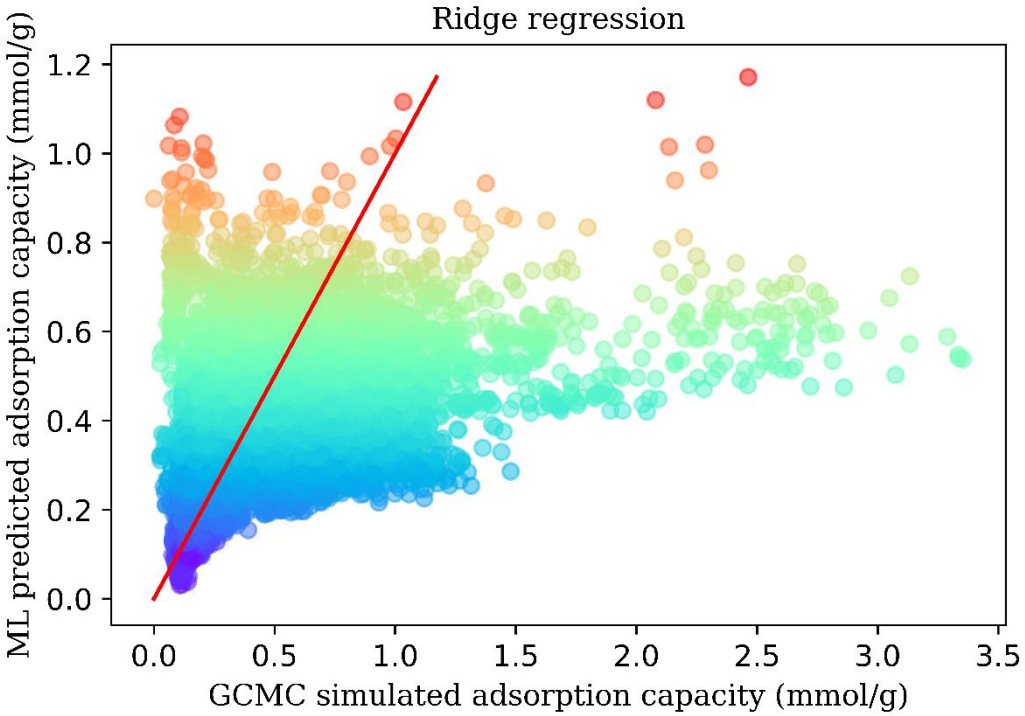


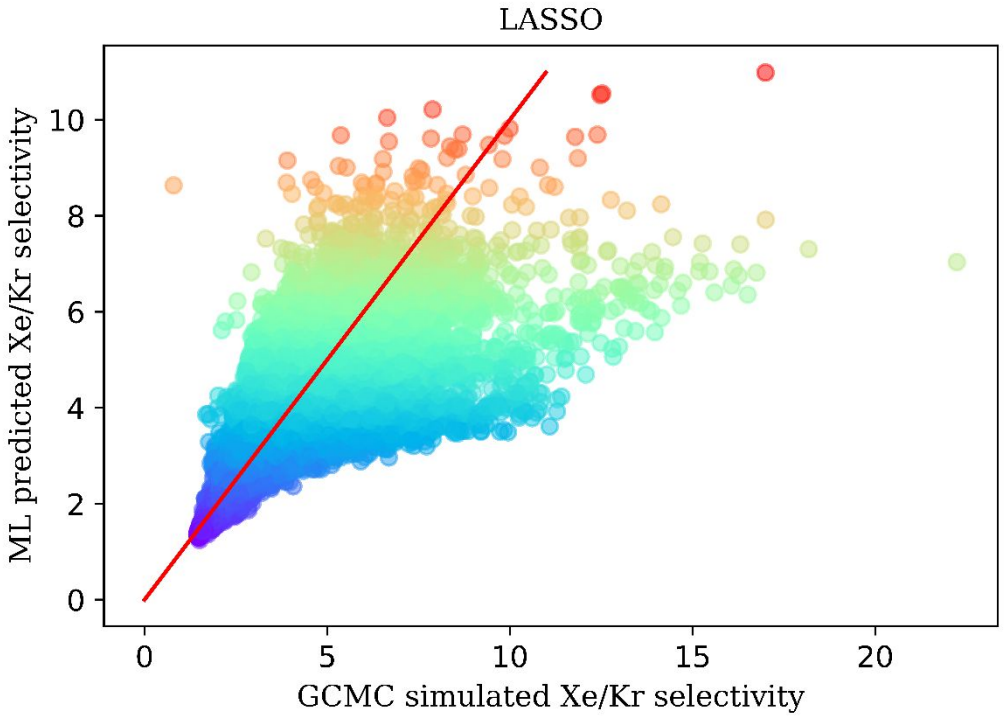
Figure S2. The parity plots between Xe/Kr selectivity and Xe uptake based on the G-MOFs database.


Ridge regression
ML predicted Xe/Kr selectivity
GCMC simulated Xe/Kr selectivity
0
2
4
6
8
10
0
5
10
15
20

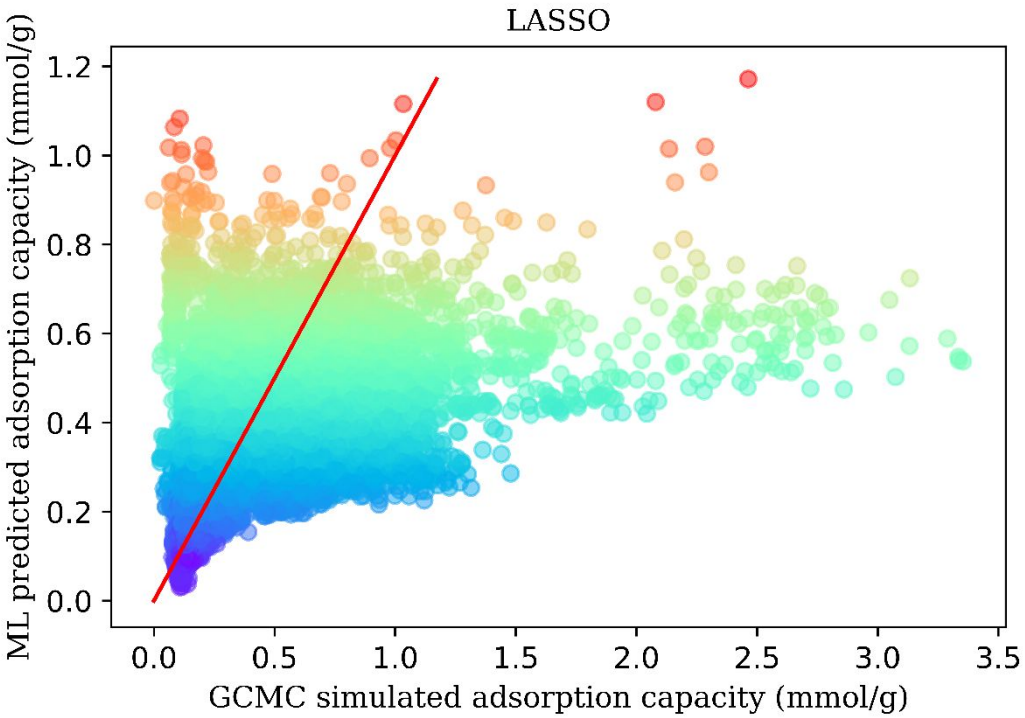
(a-1)


Ridge regression
ML predicted adsorption capacity (mmol/g)
GCMC simulated adsorption capacity (mmol/g)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

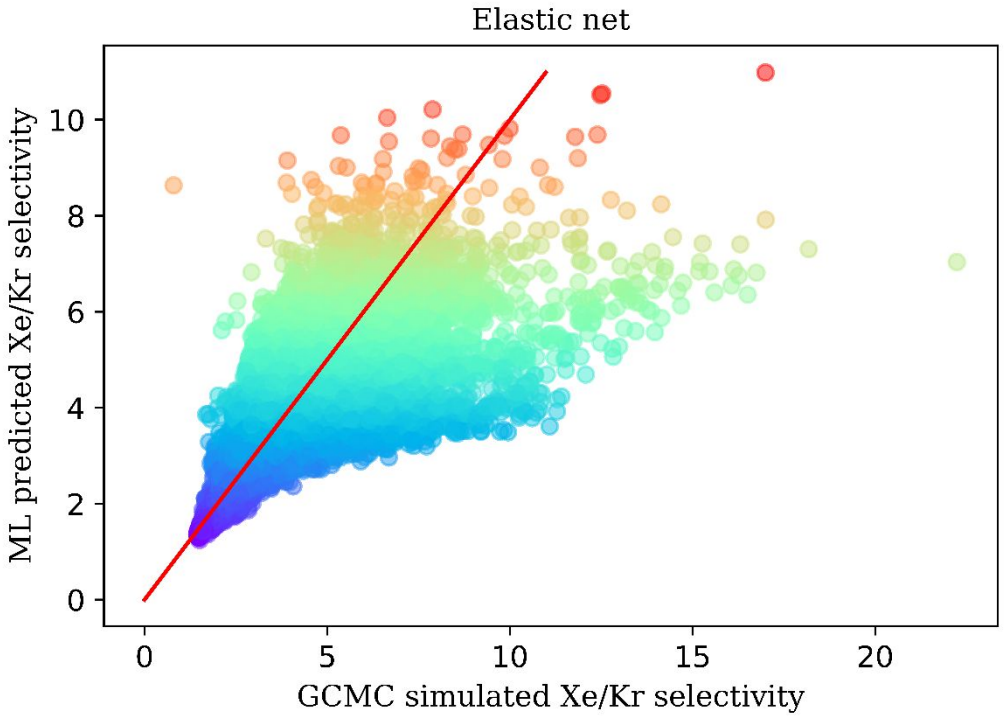
(a-2)


LASSO
ML predicted Xe/Kr selectivity
GCMC simulated Xe/Kr selectivity
0
2
4
6
8
10
0
5
10
15
20

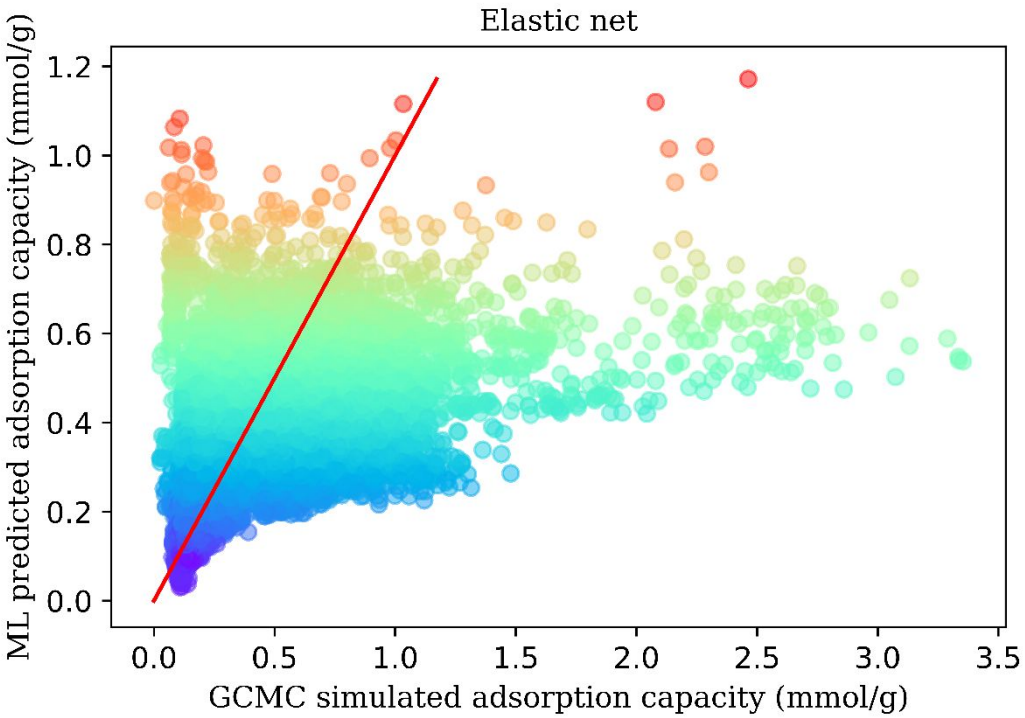
(b-1)


LASSO
ML predicted adsorption capacity (mmol/g)
GCMC simulated adsorption capacity (mmol/g)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

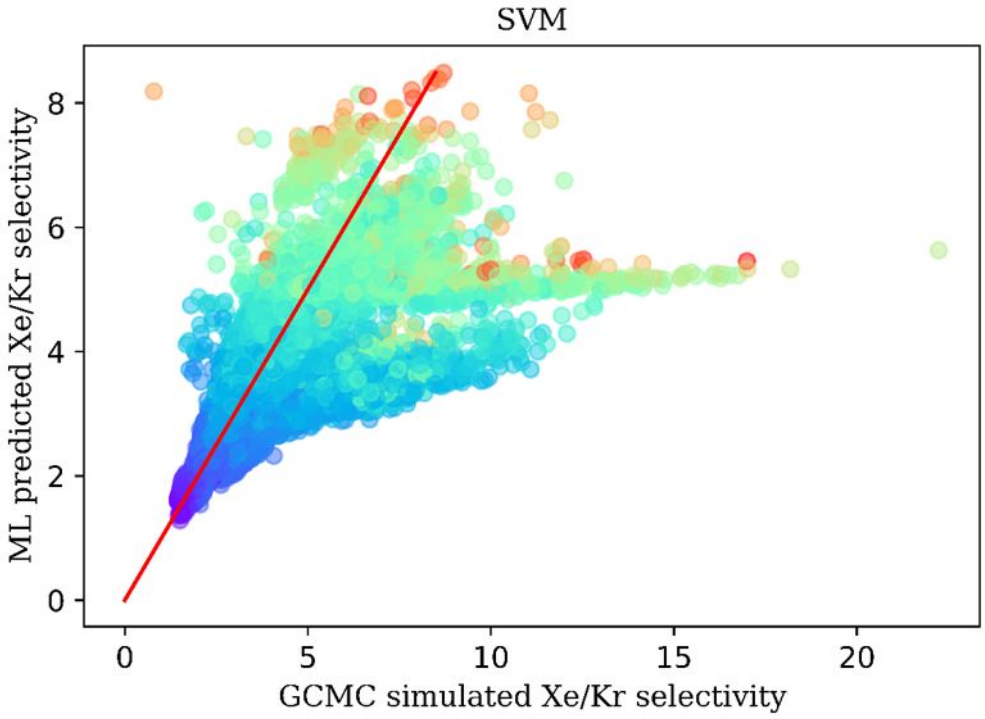
(b-2)


Elastic net
ML predicted Xe/Kr selectivity
GCMC simulated Xe/Kr selectivity
0
2
4
6
8
10
0
5
10
15
20

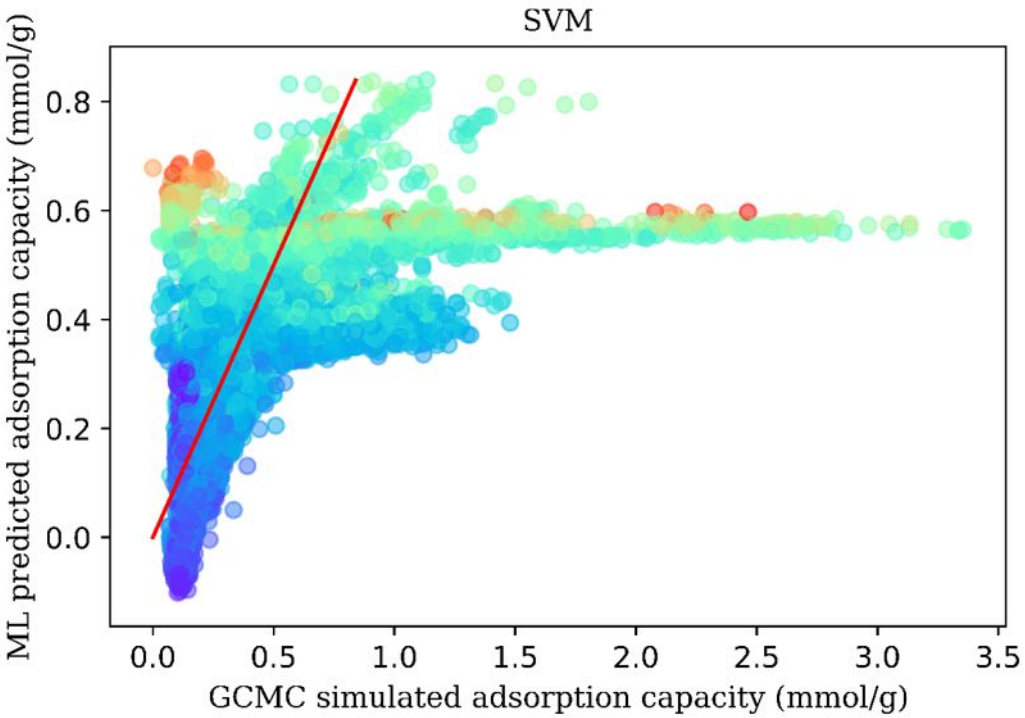
(c-1)


Elastic net
ML predicted adsorption capacity (mmol/g)
GCMC simulated adsorption capacity (mmol/g)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

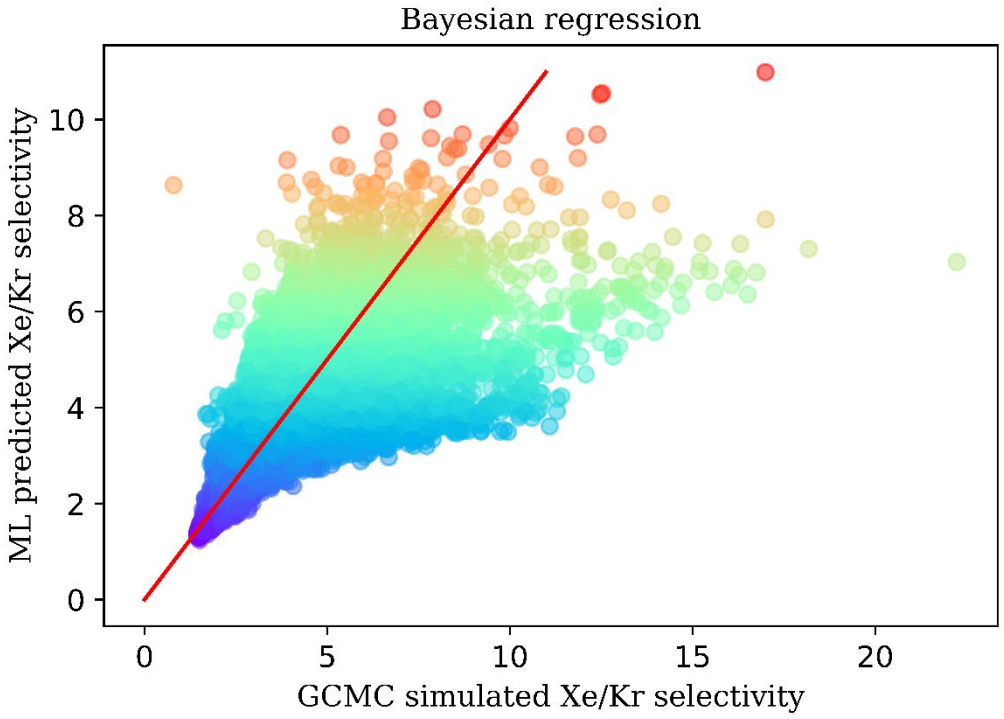
(c-2)


SVM
ML predicted Xe/Kr selectivity
GCMC simulated Xe/Kr selectivity
0
2
4
6
8
0
5
10
15
20

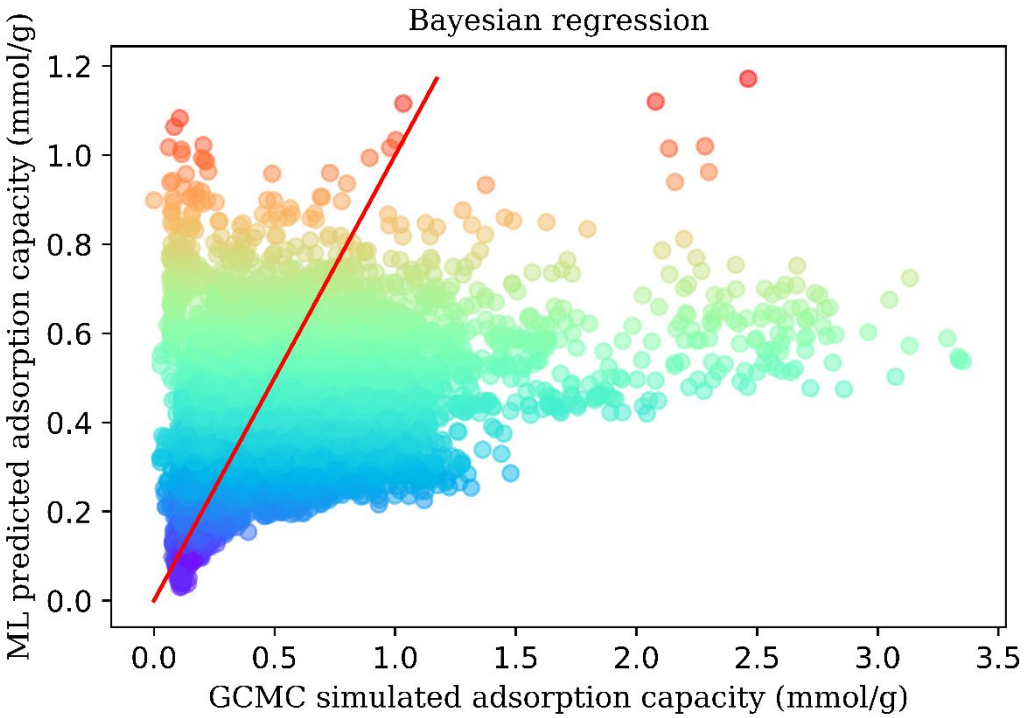
(d-1)


SVM
ML predicted adsorption capacity (mmol/g)
GCMC simulated adsorption capacity (mmol/g)
0.0
0.2
0.4
0.6
0.8
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

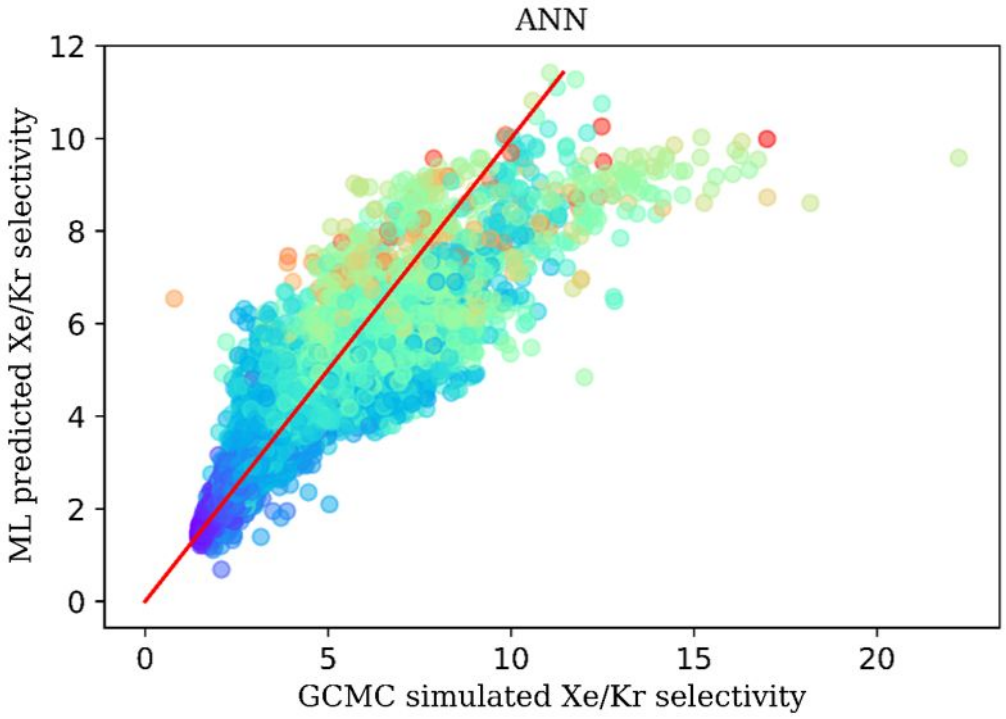
(d-2)


Bayesian regression
ML predicted Xe/Kr selectivity
GCMC simulated Xe/Kr selectivity
0
2
4
6
8
10
0
5
10
15
20

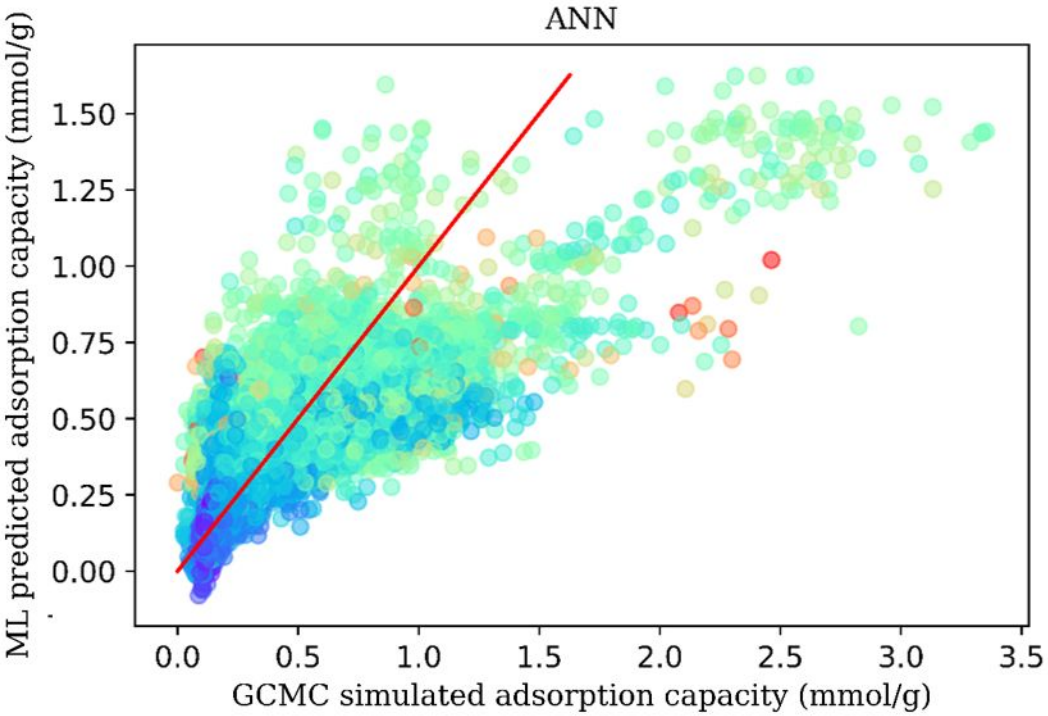
(e-1)


Bayesian regression
ML predicted adsorption capacity (mmol/g)
GCMC simulated adsorption capacity (mmol/g)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

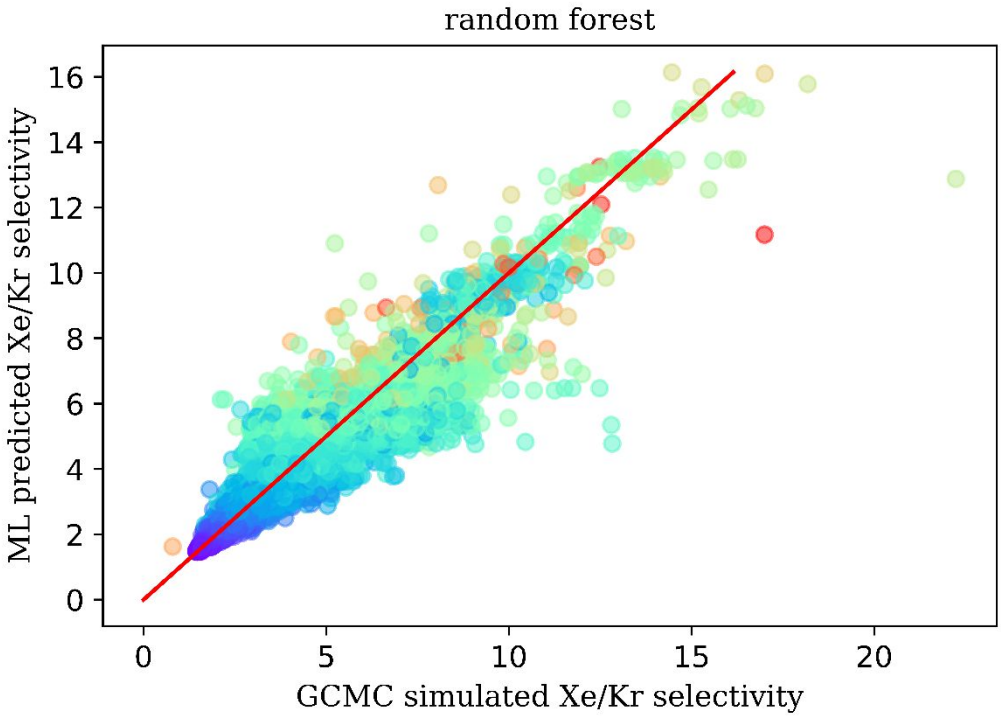
(e-2)


ANN
ML predicted Xe/Kr selectivity
GCMC simulated Xe/Kr selectivity
0
2
4
6
8
10
12
0
5
10
15
20

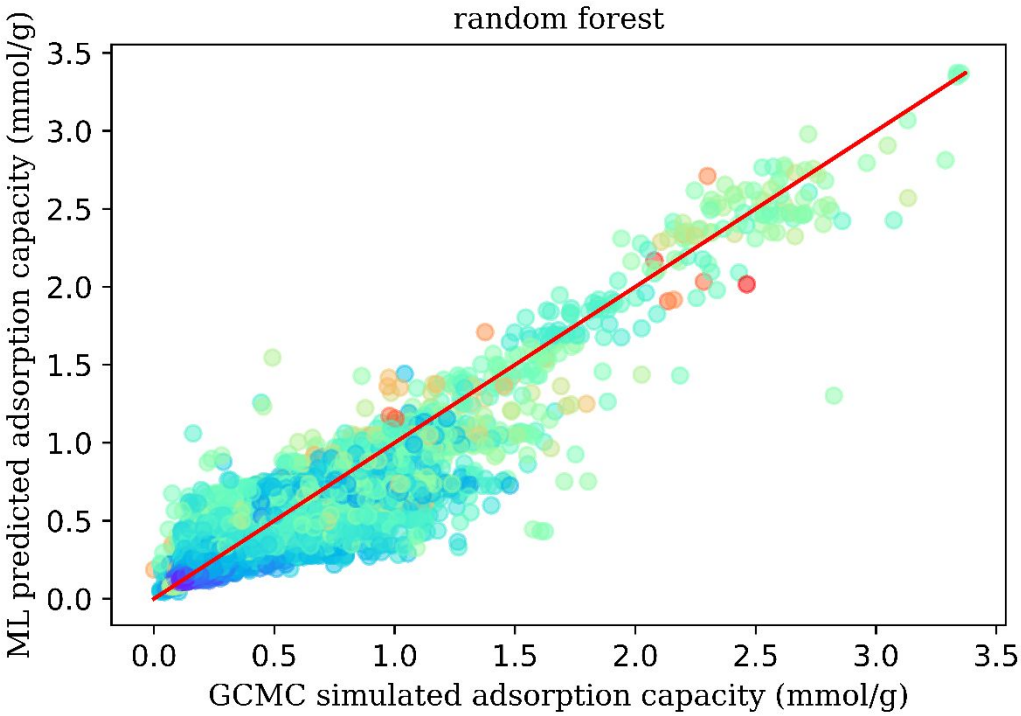
(f-1)


ANN
ML predicted adsorption capacity (mmol/g)
GCMC simulated adsorption capacity (mmol/g)
0.00
0.25
0.50
0.75
1.00
1.25
1.50
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

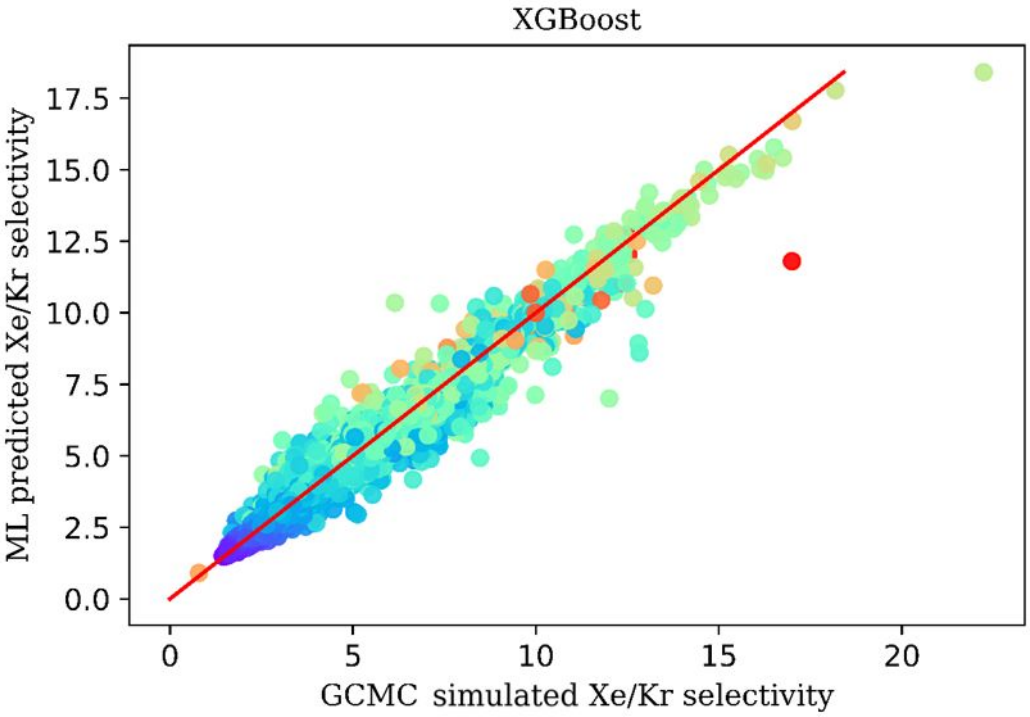
(f-2)


random forest
ML predicted Xe/Kr selectivity
GCMC simulated Xe/Kr selectivity
0
2
4
6
8
10
12
14
16
0
5
10
15
20

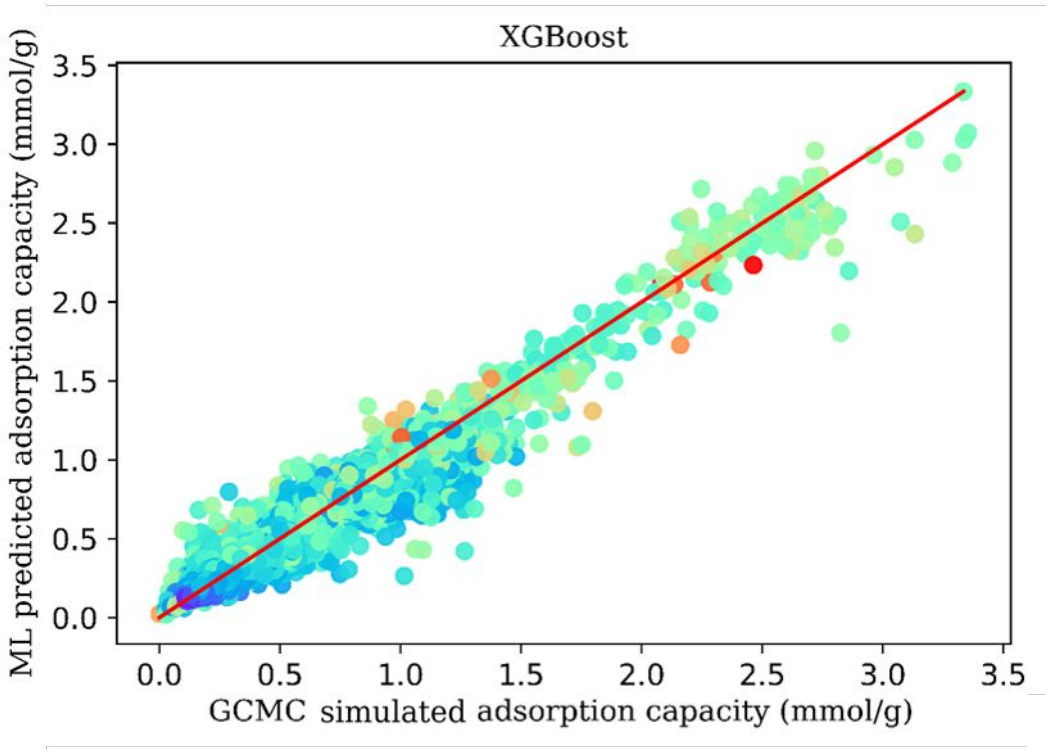
(g-1)


random forest
ML predicted adsorption capacity (mmol/g)
GCMC simulated adsorption capacity (mmol/g)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

(g-2)



(h-1)



(h-2)

Figure S3. The parity plots of Xe/Kr selectivities and Xe uptakes between GCMC simulations and different ML models prediction on the testing set based on eight

different machine learning models, including: a-1) the predicted Xe/Kr selectivities under ridge regression model. Note that red represents high density, blue represents low density; a-2) the predicted Xe uptakes under ridge regression model; b-1) the predicted Xe/Kr selectivities under LASSO model; b-2) the predicted Xe uptakes under LASSO model; c-1) the predicted of Xe/Kr selectivities under Elastic Net model; c-2) the predicted Xe uptakes under Elastic Net model; d-1) the predicted Xe/Kr selectivities under Support Vector Machine model; d-2) the prediction of Xe uptakes under Support Vector Machine model; e-1) the predicted of Xe/Kr selectivities under Bayesian Regression model; e-2) the predicted of Xe uptakes under Bayesian Regression model; f-1) the predicted of Xe/Kr selectivities under ANN model; f-2) the predicted Xe uptakes under ANN model; g-1) the predicted Xe/Kr selectivity under RF model; g-2) the predicted Xe uptakes under RF model; h-1) the predicted Xe/Kr selectivities under XGBoost model and h-2) the predicted Xe uptakes under XGBoost model. Note that each dot represents one MOF structure from the G-MOFs database. The red line represents that the GCMC calculations is same as the ML prediction results. The point closer to the red line, the more accurate adsorption property that the model predicts.

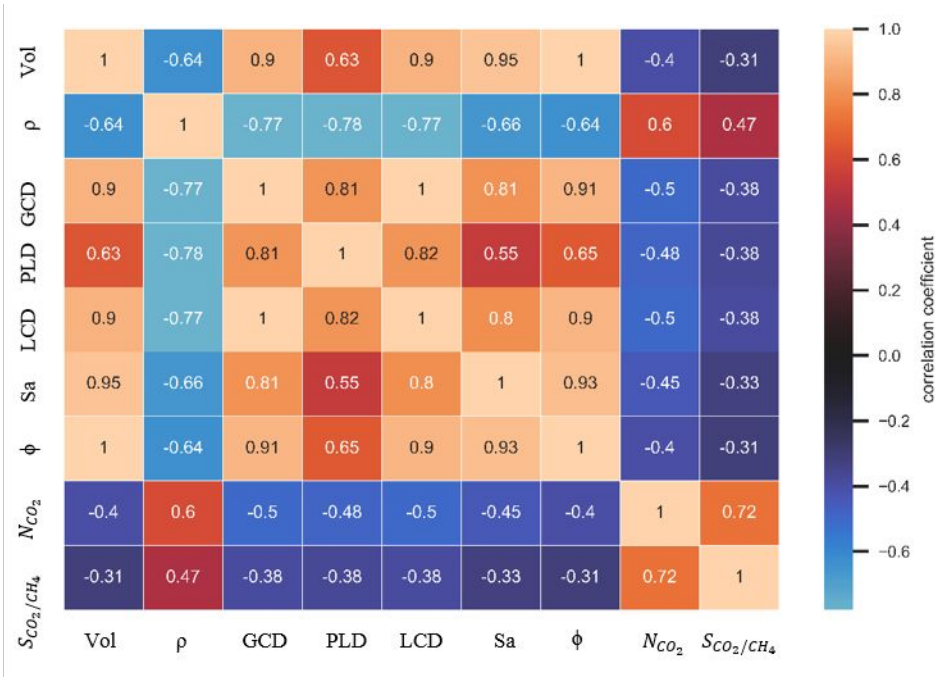
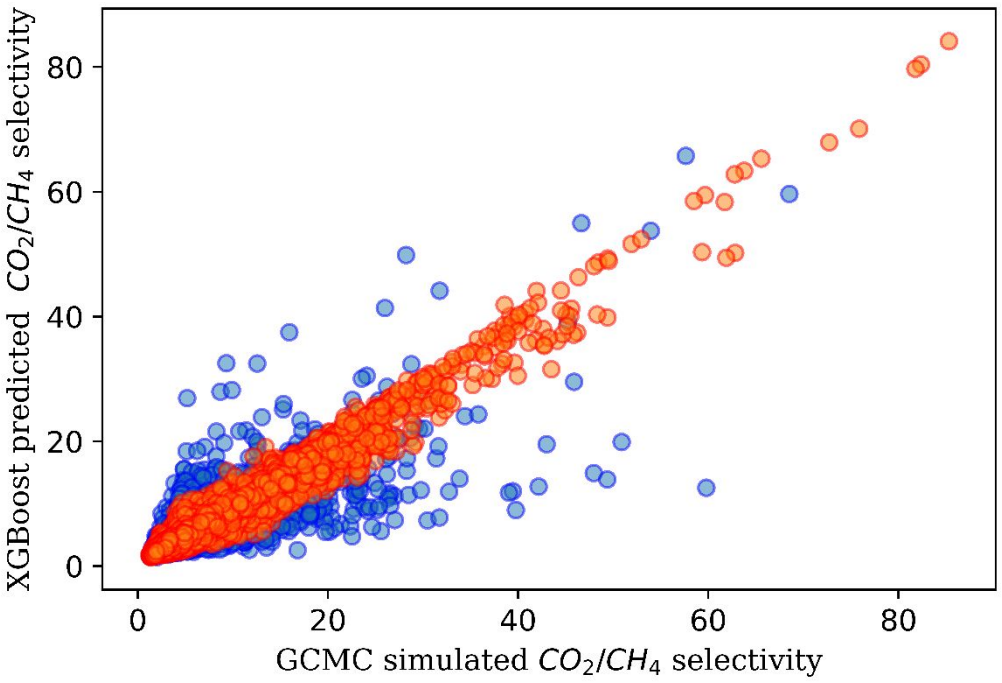
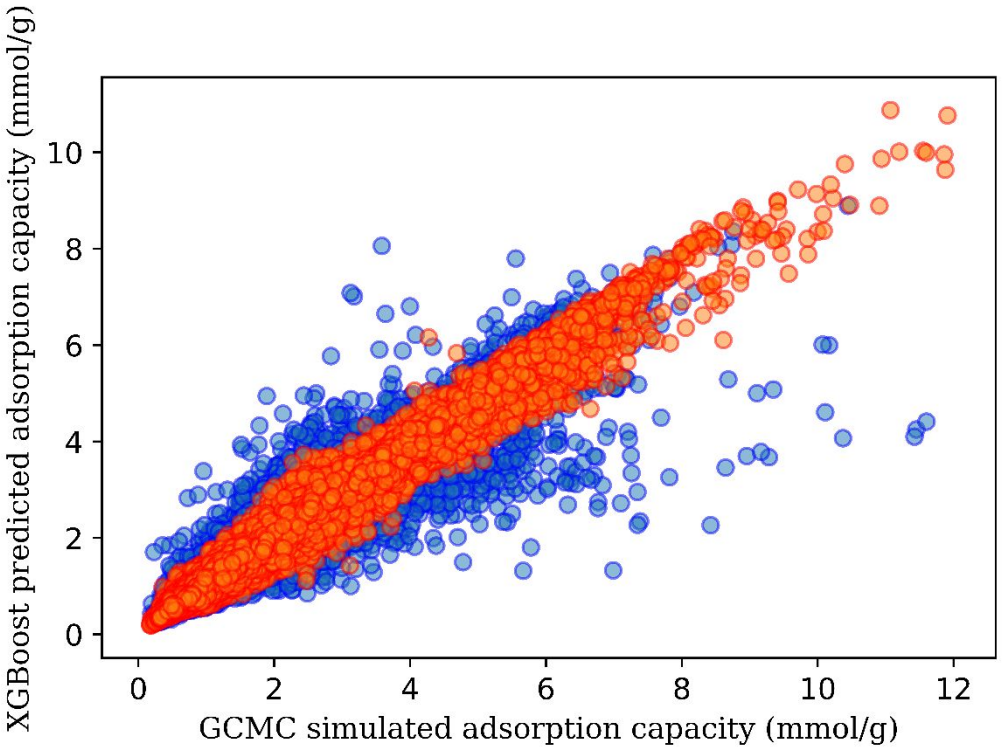


Figure S4. The plotted correlation diagram between material features and adsorption properties of CO_2/CH_4 based on G-MOFs database. Note that the color bars represent the size of the Pearson correlation coefficients.



(a)



(b)

Figure S5. The parity plots for training and testing sets data from the G-MOFs database using XGBoost model for the a) CO_2/CH_4 selectivity and b) CO_2 uptake at 1 bar and

298 K. Each dot represents one MOF structure from the G-MOFs database. The red and blue dots represent the training set and testing set data, respectively.